UiO : University of Oslo

Andrey Kutuzov

Distributional word embeddings in modeling diachronic semantic change

Thesis submitted for the degree of Philosophiae Doctor

Department of Informatics
The Faculty of Mathematics and Natural Sciences

Language Technology Group

UNIVERSITAS OSLOENSIS
MDCCCXI

2020

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo
No. 2322

ISSN 1501-7710

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

‘All perceivable form is made from this quicksilver stuff.
We call it language.’
Thoth, as retold by Alan Moore

Abstract

Words of human languages change their meaning over time. This linguistic phenomenon is known as ‘diachronic semantic change’. Such shifts are of interest both for linguists and for NLP practitioners. One possible solution for automatic large-scale modeling of semantic change is using the distributional signal. Distributional semantic models based on dense vector representations (word embeddings) are trained on large text collections and efficiently capture many aspects of word meaning. As such, they are among the foundational bricks in the building of natural language processing systems which are aimed at understanding and generating human language.

If word embeddings capture word meaning at a given point in time, then these meaning representations at different time points can naturally be compared. Diachronic word embeddings are trained on text created in different time periods. The time of creation obviously influences typical usage of words and reflects significant changes in all aspects of their meaning.

This unsupervised ‘data-driven’ detection of temporal semantic change is the main topic of the present thesis. Overall, we study what information about diachronic semantic processes is captured by distributional vector representations. We train diachronic embeddings in different ways, and devise methods which use them to solve the task of detecting how words change their meaning and usage over time.

In particular, we first survey and systematize previous work on the topic, including ours. Then, we successfully conduct cross-lingual analysis of the speed of semantic change in evaluative adjectives. We propose novel ways of evaluation for semantic change detection methods based on word embeddings. In particular, it is described how the dynamics of real-world events like armed conflicts is reflected in the changes which temporally-aware distributional representations undergo. This allows manually annotated armed conflict datasets to function as a proxy gold standard to evaluate semantic change detection methods and probe diachronic word embeddings for their temporal awareness. We show that this holds not only for single words, but also for typed semantic relations between them as well.

Finally, we evaluate the potential of contextualized word embedding architectures like BERT and ELMo for modeling diachronic semantic change. We show that they outperform the methods based on traditional ‘static’ embeddings, while providing richer possibilities for visualization and qualitative analysis. At the same time, we identify and categorize possible issues which a historical linguist might encounter when using contextualized architectures in an attempt to trace diachronic semantic shifts.

Acknowledgments

The thesis you are going (or not) to read would not have been possible without input from many people. First and foremost, I am thankful to my supervisors Erik Velldal and Lilja Øvrelid. Throughout these five years, they were always ready to answer my (often silly) questions and to guide me on the thorny way of a PhD candidate. It is very important to have someone to learn something from: Erik and Lilja did their part of this semantic frame perfectly (not sure about me, though).

But Erik and Lilja were not alone. The Language Technology Group at the University of Oslo is full of wonderful people, and I really enjoyed every day here. Arne, Eivind, Elizabeth, Eman, Farhad, Ildikó, Jan Tore, Jeremy, Maja, Milen, Murhaf, Petter, Pierre, Samia, Stephan, Taraka, Vinit, all of you contributed to my thesis in some way, even if unknowingly. Thank you, colleagues and friends. May the coffee never end.

It is impossible not to mention the School of Linguistics at the Higher School of Economics in Moscow, where I got my Master degree in computational linguistics. The concentration of bright minds and bright eyes there breaks any measuring device, whether we are talking about staff or about students. I would have never become what I am without the HSE. Considering the topic of this thesis, it is especially important to thank the HSE students who worked in the diachronic embeddings research group supervised by me: Daria Bakshandaeva, Vadim Fomin, Vladislav Mikhailov and Julia Rodina. Folks, you are awesome.

I am grateful to Chris Biemann and Alexander Panchenko for hosting me at their lab during my stay in Hamburg in 2018. Thrilling discussions and the whole research experience in Hamburg made me much more confident in my NLP skills. Hope to collaborate with you again in the future!

Big thanks go to my long-standing co-authors Maria Kunilovskaya and Elizaveta Kuzmenko. Sorry for being grumpy sometimes. I know I can always come to you with another weird idea: be sure this is mutual.

This thesis (not to mention its author) would not be possible without my mother. After all, who was bringing all these books for me to read all the time? I am immensely thankful for that.

Daria, your love and support is incredible. I will say more off-the-record.

Finally, I thank all those who shine through the darkness all around the world.

Andrey Kutuzov
Oslo, October 2020

Contents

Abstract iii

Acknowledgments v

Contents vii

List of Figures ix

List of Tables xiii

1 Introduction 1
1.1 Terminological issues 4
1.2 Research questions 7
1.3 Thesis outline 10
1.4 Publications 12

2 Distributional modeling of meaning 13
2.1 The distributional hypothesis 13
2.2 Working with co-occurrence matrices 16
2.3 The curse of dimensionality 19
2.4 Rise of machine-learned distributional models 20
2.5 Shortcomings of word embeddings 23
2.6 Recent trends in distributional semantic modeling 30
2.7 Other approaches to model meaning 31
2.8 Summary 33

3 Modeling diachronic semantic change: state of the field 35
3.1 Semantic shift as a linguistic concept 36
3.2 Tracing semantic shifts distributionally 39
3.3 Laws of semantic change 56
3.4 Diachronic semantic relations 58
3.5 Applications 61
3.6 Summary 62

4 Measuring diachronic evolution of evaluative adjectives 65
4.1 Motivation 65
4.2 Training corpora and evaluative lexicons 67
4.3 Estimating the speed of semantic change 71
4.4 Experimental results 73
4.5 Limitations 78

4.6 Summary 79

5 Semantic change and world events: armed conflict dynamics 81

5.1 Armed conflict research data 82

5.2 From nearest neighbors to anchor words: tracing armed conflicts 89

5.3 Tracing shifts in semantic relations 101

5.4 Forecasting future armed conflicts as diachronic one-to-X analogies 116

5.5 Summary 124

6 Contextualized embeddings and semantic change 127

6.1 Embeddings from Language Models (ELMo) as contextualizers 128

6.2 Ways of comparing contextualized embeddings over time . 131

6.3 Measuring lexical ambiguity with contextualized token embeddings 136

6.4 Empirical evaluation of contextualized methods 149

6.5 Qualitative analysis 168

6.6 Summary 182

7 Conclusion 185

7.1 General contributions 187

7.2 Summary 188

7.3 Future work 195

7.4 Publicly available code, datasets and models 197

Appendices 199

A Structure of the GED dataset 201

B Per-year accuracies for next-year armed group predictions 205

Bibliography 209

List of Figures

2.1 Contexts of the word ‘*tea*’ 14
2.2 Contexts of the word ‘*coffee*’ 14
2.3 The nearest neighbors of the English noun ‘*shift*’ in the distributional model trained on the English Wikipedia. 17
2.4 Two-dimensional t-SNE projection of 10 word vectors from the distributional model trained on the Google News corpus. The colors were added manually to easily distinguish the lexical groups containing 0) capitals, 1) countries, 2) common nouns. 20
2.5 The now-famous depiction of the two *word2vec* architectures (CBOW ans Skip-gram) from Mikolov, K. Chen, et al. (2013). . 22
2.6 Interplay between gold human judgments and distributional model predictions on lexical semantic similarities. 28

3.1 Tensor representation of a semantic space; image from (Jurgens and Stevens, 2009). 46
3.2 Semantic trajectory of the English word ‘*gay*’ in the space of its context words (Kulkarni et al., 2015). 50
3.3 Incremental training of word embeddings (in this example, CBOW models are trained on yearly diachronic corpora). 54
3.4 Analogy task in a vector space, with the results from two distributional models (WebVectors service (Fares et al., 2017)). . 59
3.5 Diachronic changes of emotions associated with the English word ‘*climate*’ as visualized by JeSeMe (Hellrich et al., 2018). The vertical axis (labels omitted for readability) reflects the values of emotion scales, with 0 in the middle. 62
3.6 Modeling diachronic semantic change distributionally: research timeline 63

4.1 Alterations in the nearest distributional neighbors of the English adjective ‘*monumental*’: from ‘*sculpture*’ in the 1960s to ‘*awesome*’ in the 2000s. t-SNE projection of CBOW vectors trained on the COHA corpus. 67
4.2 Alterations in meaning of the Russian adjective ‘бескомпромиссный’ (‘*uncompromising*’): from ‘беспощадный’ ‘*ruthless*’ over ‘фанатический’ ‘*fanatical*’, ‘страстность’ ‘*passion*’, to ‘убежденность’ ‘*conviction*’, ‘героика’ ‘*heroic*’ to ‘непримиримость’ ‘*intransigence*’, ‘противостояние’ ‘*confrontation*’. English equivalents are given in red. 73

5.1 Number of armed conflicts in the world per year. Colors denote conflict types: pink stands for one-sided violence, scarlet red stands for state-based violence, and dark red stands for non-state violence (source: Uppsala Conflict Data Program). 83

5.2 A word embedding model trained on news texts (Gigaword) correctly answering a synchronic armed conflict analogy question (predicting the ULFA armed group active in India). 105

5.3 Locations (red) and corresponding armed groups (blue) active in 2000; t-SNE projection of high-dimensional word embeddings. . 111

5.4 Locations (red) and corresponding armed groups (blue) active in 2001; t-SNE projection of high-dimensional word embeddings. . 112

5.5 Locations (red) and corresponding armed groups (blue) active in 2002; t-SNE projection of high-dimensional word embeddings. . 113

5.6 Prediction of armed groups active in Algeria in 2014, based on a transformation matrix (red arrow) learned from the 2013 data; 2-dimensional PCA projection. 122

5.7 Prediction of armed groups active in Yemen in 2011, based on a transformation matrix (red arrow) learned from the 2010 data; 2-dimensional PCA projection. 123

6.1 ELMo architecture. Image by Karan Purohit, https://medium.com/saarthi-ai/elmo-for-contextual-word-embedding-for-text-classification-24c9693b0045 129

6.2 Correlation between ambiguity metrics of a word and the number of its WordNet synsets given different frequency thresholds. . . . 141

6.3 Median word frequencies (log scale) for the intersection of COHA sub-corpora vocabularies. Red vertical lines stand for our frequency tier boundaries. 144

6.4 Dynamics of ELMo lexical ambiguity (DIV). Left: pre-trained on English Wikipedia; center: pre-trained on the COHA corpus; right: trained incrementally on the COHA corpus. 145

6.5 Changes in ELMo diversity brought by each decade. Left: Wikipedia model; center: single COHA model; right: corresponding incremental models. 146

6.6 Changes in ambiguity for one and the same decade – 1980s – as calculated with five different incremental ELMo models. Left: mid frequency words; right: low frequency words. 148

6.7 Target word frequencies in the semantic change test sets we use (based on the respective Wikipedia corpus for each language). Five German words and one Swedish word omitted from the plot for the sake of visual convenience, since their frequency ranks are below even 400 000. 152

6.8 Top: distribution of semantic change degree in the **gold** data; middle: distribution of scores predicted by the **APD** method; bottom: distribution of scores predicted by the **PRT** method. . 166

6.9 Performance of the PRT and APD contextualized methods depending on the median gold score. 167
6.10 PCA projections of ELMo representations of each occurrence of the word '*cell*' in four different decades: actual semantic shift. . 169
6.11 PCA projection of ELMo token representations of each occurrence of the word '*cell*' in the 2000s, with clusters labeled with senses. 171
6.12 PCA projection of ELMo contextualized representations of occurrences of the word '*cell*' in all five COHA decades (single COHA model). 172
6.13 PCA projection of ELMo contextualized representations of occurrences of the word '*virtual*' in all five COHA decades (single COHA model). 174
6.14 PCA projections of ELMo token embeddings for '*banish*', '*mg*', '*progressive*' and '*indirect*' across all five COHA decades. 177
6.15 PCA projections of ELMo representations of each occurrence of the word '*phone*' in four different decades: stable syntactic clusters. 181

List of Tables

2.1 A simple example of a co-occurrence matrix. 16

2.2 Example of the SimLex999 semantic similarity test set from Hill et al. (2015), along with predictions from a distributional model (the 'output' column). 26

3.1 Popular English corpora for diachronic research 42

4.1 Corpora sizes (in millions of words). 68

4.2 Evaluative adjective lexicons. 70

4.3 Differences in the intensity of semantic change between evaluative adjectives and fillers. Positive values correspond to evaluatives changing significantly faster, and vice versa. 74

4.4 Correlation of semantic change speed and normalized word frequency across all adjectives (evaluative and fillers). Positive values correspond to frequent words changing significantly faster, and vice versa. 75

4.5 Difference in the intensity of semantic change between evaluative adjectives and fillers (frequency > 100). Positive values correspond to evaluatives changing significantly faster ($p < 0.1$), and vice versa. 76

5.1 Macro-F1 scores for predicting conflict state changes (ternary classification). 97

5.2 Detailed performance of the best armed conflict state change detection method (results of weighted random guess in parentheses). 99

5.3 Accuracies at different number of nearest neighbors (k) for synchronic projections (time periods not taken into account) from locations to armed groups, and vice versa. The Vector Offset baseline is also reported. 107

5.4 Accuracies at different number of nearest neighbors (k) for synchronic projections on the pairs from semantic sections of the Google Analogies test set ($\lambda = 1.0$). 108

5.5 Accuracies of 2000 → 2001 diachronic projection in armed group detection at different number of nearest neighbors (k). 111

5.6 Average accuracies of predicting next-year insurgents based on location vectors and projections learned from the previous year. 115

5.7 Comparative statistics of our armed conflict test sets. 119

5.8 Average recall of diachronic analogy inference 120

5.9 Average diachronic performance of armed conflicts prediction with cosine thresholding: testing on conflict areas only. 122

5.10 Average diachronic performance of armed conflicts prediction with cosine thresholding: testing on all areas. 123
5.11 Average synchronic performance of armed conflicts prediction with cosine thresholding: testing on all areas. 124

6.1 The Spearman rank correlations between the number of WordNet synsets and ambiguity scores or word frequency: all words. . . . 139
6.2 The Spearman rank correlations between the number of WordNet synsets and ambiguity scores or word frequency: only words with frequency higher than 130. 140
6.3 Quantitative characteristics of the lexical semantic change test corpora (w denotes target words). Slashes separate counts for the older and the newer sub-corpus in each pair. 151
6.4 The correlations between our predictions and human judgments from the GEMS test set (change between the 1960s and the 1990s). 155
6.5 Results for SemEval-2020 Task 1 sub-task 2 on English: Spearman correlation. Scores for BERT-Large are given in parentheses. . . 157
6.6 Results for SemEval-2020 Task 1 sub-task 2 on German: Spearman correlation. 159
6.7 Results for SemEval-2020 Task 1 sub-task 2 on Latin: Spearman correlation. 160
6.8 Results for SemEval-2020 Task 1 sub-task 2 on Swedish: Spearman correlation. 161
6.9 Correlation scores for the methods under analysis on SemEval-2020 Task 1 Subtask 2 averaged over four languages. The number of † denotes the number of languages for which the correlation was statistically significant ($p < 0.05$). 163
6.10 Spearman correlations per test set for our best methods. * denotes statistically significant correlation ($p < 0.05$). 164
6.11 PRT/APD predicted semantic change scores and their frequency differences (δ) across five COHA decades for the words '*cell*' and '*virtual*'. IPM stands for 'instances per million'. 174
6.12 10 points of the strongest change in five decades of COHA (as measured by PRT/APD). Word color indicates its class. 176

B.1 Yearly accuracies (including OOV words) of predicting next-year insurgents based on location vectors and projections learned from the previous year. Best approach from Section 5.3. 206
B.2 Yearly F1 scores of armed conflicts prediction in the One-to-X setup (Gigaword dataset). 207
B.3 Yearly F1 scores of armed conflicts prediction in the One-to-X setup (NoW dataset). 207

Chapter 1

Introduction

This thesis explores the use of distributional word embeddings to model various types of diachronic semantic change.

The natural language processing (NLP) community has in last decade seen the development of highly efficient and effective representations of natural language semantics: so-called *word embeddings* (dense vector representations of meaning). Conceptually, they are strongly related to the 'distributional hypothesis' (Firth, 1957), assuming that word meaning can be approximated by patterns of word usage (unlike other approaches to meaning like correspondence theory or formal semantics). This is also how we understand the notion of 'lexical semantics' or 'word meaning' within the frame of this thesis. This approach is relatively straightforward to formalize, and it can leverage the large troves of textual data available nowadays.

Word embedding architectures used today constitute the culmination of decades of work on various other forms of vector-based representations of distributional information from text, starting at least from Jones (1972) or even earlier. The differences between dense word embeddings and other distributional approaches are extensively covered in Chapter 2. It suffices to say here that they serve as foundational building blocks of many modern NLP systems, by providing convenient semantically-aware lexical input representations to architectures based on artificial neural networks. State-of-the-art approaches to construct embeddings themselves often also employ neural networks of varying depth and complexity (Bengio et al., 2003; Mikolov, Yih, et al., 2013). Recent advances in this field include the advent of 'contextualized' models which produce context-sensitive representations, in contrast to traditional 'static' embeddings (Melamud et al., 2016; McCann et al., 2017; Peters, Neumann, Iyyer, et al., 2018). We discuss and employ contextualized architectures in Chapter 6.

The ability of word embeddings and related methods to capture many aspects of the semantics of human languages for multiple practical tasks is now widely acknowledged (Baroni et al., 2014; Goldberg, 2017; Desagulier, 2017; J. Johnson et al., 2019).[1] An important field of research is general intrinsic evaluation of such models: to reliably evaluate the ability of distributional systems to capture meaning, one has to move beyond the limits of a purely distributional understanding of semantics. This is the only way to avoid the circularity of postulating that word usage does reflect word usage. For this reason, distributional representations are often evaluated using knowledge about word meaning coming from other sources (for example, annotation by human

[1] As a rule, throughout the thesis we use the authors' last names in citations. In rare cases when more than one different authors share the same last name in our bibliography, we differentiate them by adding initials ('J. Johnson').

informants). We also rely on this approach in several of the chapters below.

In all languages, words' meanings and usages evolve as time passes by. Other types of semantic change can also be studied: for example, cross-domain shifts, when a word is regularly used in a different sense in a particular genre or domain. But in this thesis, we restrict ourselves to sequential diachronic changes that words undergo over time. Note also that the processes of grammaticalization (in which a word loses its lexical meaning and becomes a grammatical form) fall outside the scope of the present thesis, although in theory they also can be seen as diachronic semantic change.

Diachronic changes in word meaning have been studied by linguists for a long time (Bréal, 1899; Stern, 1931; Bloomfield, 1933): they focused on many different aspects of word meaning, including, but not limited to the notion of 'word senses'. Multiple classifications of semantic change have been developed; we discuss them in detail in Chapter 3. Bloomfield (1933) was, arguably, the first to coin the term 'semantic shift', which has been used extensively since then. In Section 1.1 below we explain our understanding of 'semantic shifts'. It is important to note here that we use 'diachronic semantic change' as an umbrella term potentially covering all aspects of meaning; throughout the thesis, we specify what particular kind of change is meant, when necessary.

If one takes the distributional perspective on meaning (see Chapter 2), word co-occurrence data should be enough to detect semantic change. As already mentioned, there exists an extensive ecosystem of benchmark tests to evaluate the performance of semantic representations. Evaluation results have shown that word embeddings capture many aspects of *synchronic* lexical semantics quite well. Hence, they are natural candidates to employ also for modeling semantic change in a *diachronic* setup, which is the topic of this thesis. We fill in the lack of analysis on temporal abilities of word embeddings and study what information about diachronic semantic change in natural languages can be captured by such architectures, and how this information can be extracted from them most efficiently. Thus, we naturally deal with the analysis of two subjects:

1. *Diachronic word embeddings*, that is distributional representations inferred from time-specific collections of texts. We test and evaluate different methods of creating such representations.

2. *Embedding-based computational semantic change detection systems.* This subject comprises creating and evaluating possible approaches and algorithms which employ word embeddings (diachronic or not) to model semantic change.

The second subject is the most important, since pre-trained representations alone do not allow one to actually model, analyze or detect semantic change. One has to come up with proper algorithms to extract temporal semantic information captured in word embeddings (see Chapter 3 for more details).

This thesis can be also be looked at as an attempt at *probing* word representations for diachronic information. Since word embeddings are technically just dense matrices of float values, they are not directly interpretable by humans, which relates to what is often dubbed the 'blackbox problem' of neural networks (Linzen et al., 2019). A large volume of research is devoted to probing embeddings or other neural NLP systems (using a multitude of methods) to find out what information they encode, or what they 'know' about language (Yaghoobzadeh, Kann, et al., 2019). One can probe embeddings for many different aspects of linguistic knowledge. This thesis explores their ability to capture *language change*: more specifically, temporal changes in various aspects of lexical semantics, including semantic relations between words.

Notably, there exist several different families of computational approaches which allow for the extraction of information related to semantic change from corpus data, starting from frequency analysis and down to more complex methods, including those based on word embeddings (but not limited to them). It is next to impossible to cover *all* possible approaches, although we try to discuss at least the most important ones in Chapter 3. Here we emphasize again that this thesis is focused on the methods which employ distributional dense embeddings of various types (note that word embeddings themselves of course do not constitute a semantic change detection method).

None of the existing approaches (including ours) are entirely satisfactory in their performance or coverage. One of the reasons for that is that the notion of diachronic semantic change modeling itself is open to multiple interpretations, and different methods can be successful in different aspects of it. Distributional modeling is not a silver bullet for any diachronic linguistic task. However, the current state of affairs in NLP makes dense word embeddings inferred from distributional information the first-order candidates to built semantic change detection systems on: both because of their performance in various synchronic semantic tasks, and for technical reasons (they are computationally efficient and easy to integrate into deep learning architectures). We believe this makes it worthy to study and probe what kind of diachronic information word embeddings capture, including information about semantic change. It is also important to find how embedding-based methods compare to other approaches (distributional or not).

In 2015, when the work on this thesis started, the field of tracing semantic change with diachronic word embeddings was still in its early stage. Also, the word embeddings field itself was much younger: for example, contextualized architectures did not exist back then, and were not introduced until 2018. In the next years, concurrently with the development of the thesis, a significant amount of research was done and published, including by the author (expanded versions of some of these publications are included here as chapters). In particular, various types of distributional vector models have extensively been evaluated with regards to their usefulness for the various aspects of the task. The present thesis contributes to this growing field by systematically presenting the results of our experiments and proposing new ways of approaching the problem of semantic change modeling (see the description of our research questions below).

One of the important results of the growing interest in this topic from the natural language processing community was a stricter formulation of the task of diachronic semantic change modeling. In this thesis, we mostly deal with the following aspects of it:

1. *Binary classification* of linguistic entities based on whether they *have undergone a semantic change* (for example, 'has the word 'X' changed semantically in the time period 1 compared to the time period 0 or not?'). Again, the notion of 'undergoing a semantic change' can be understood in multiple ways (see Section 1.1).

2. *Estimating and quantifying the degree of semantic change* (for example, 'has the word 'X' changed semantically more or less than the word 'Y' in the time period 1 compared to the time period 0?'). This is often cast as a ranking task.

Note that an important part of lexical semantics is *relational* in its nature. By 'relational' we here mean 'dealing with the links between the meanings of different words'. This is reflected both in ontology-based computational semantics (WordNet is a graph of *semantic relations* between senses), and in the recent applications of distributional architectures. One of the most interesting properties of contemporary word embeddings is their ability to capture linguistic regularities, like the similarity of relations between ('*father*', '*son*') and ('*mother*', '*daughter*'); this gave birth to the task of analogical reasoning (Mikolov, Yih, et al., 2013). Up to now, the encoding of such semantic relations in word embedding models was studied and evaluated only synchronically. However, these relations can also shift and change over time: for example, the second item from the list above can be formulated as 'has the semantic relation between the word 'X' and the word 'Y' changed more or less than the same relation between the word 'Z' and the word 'W' in time period 1 compared to time period 0?', thus introducing a diachronic aspect. We describe various NLP tasks associated with this in Section 3.4 and apply distributional word embeddings to their modeling in Section 5.4. It should be emphasized that we consider such phenomena to still fall within the scope of diachronic semantic change studies. When focusing on changes in the meaning of individual words (i.e. without such a relational focus), the term 'lexical semantic change detection' (LSCD) is often used (Schlechtweg, Hätty, et al., 2019).

1.1 Terminological issues

The terms 'semantic shift' and 'semantic language change' can be vague (both in synchronic and diachronic contexts). We will now make clear their usage in this thesis.

Semantics of natural language is a complex phenomenon, as it describes many levels of language and involves many aspects of the interplay between signs and their meanings. In the most common understanding of the term (Bloomfield, 1933), a 'semantic shift' occurs when a word changes the set of its respective *senses*: by acquiring a new sense, losing an existing one, or both.

The definition of the term 'sense' itself can be discussed and problematized (Kilgarriff, 1997), but for the time being we will define it in a lexicographic way: as an entry in a dictionary, where one and the same word form can have several entries, corresponding to different 'senses' of this word and together forming its 'conventional meaning'.

A semantic shift of this sort occurred to the English word '*cell*' between the 1990s and 2000s (see a more detailed description in Chapter 6). The word acquired a new sense of 'MOBILE PHONE' in addition to the already existing senses of 'PRISON CELL' and 'BIOLOGICAL CELL'. Its meaning has undergone a shift (a new entry was added to its existing sense inventory). An NLP system could then be tasked with the problem of detecting or analyzing this shift and other similar shifts, based on corpora of natural language texts created in the 1990s and in the 2000s. The general idea underlying a distributional approach to this task (which is the approach adopted in this work) is that semantic shifts cause notable changes in typical word co-occurrences.[2] In the case of '*cell*', in the 2000s, this word started to frequently co-occur with context words like '*phone*', '*call*', '*ring*', etc., almost non-existent in its surroundings before the advent of cellular communication.

The whole concept of lexicographic 'senses' implies discreteness: a word either has one, two, three or k senses, each with its own dictionary definition. A shift occurs when this set changes in any way.[3] Let us define this case as *semantic shifts proper.*

However, there is more to semantic change than that. A fundamental notion in semantics concerns the different types of relations between 'senses' of one and the same word form as it occurs in natural texts (semantic proximity). Consider the well-known distinction between homonymy and polysemy (both being cases of colexification). Let us also denote n different occurrences of the word X as $X_{0...n}$. In the case of *homonymic* relation between, for example, X_0 and X_1, the senses behind these two occurrences are completely unrelated to each other. In the case of *polysemy*, the senses behind X_0 and X_1 are still different, but now they are related (often it is true for senses which appeared historically after splitting another old sense into several). In the case of a *monosemous* word, the senses behind all $X_{0...n}$ members are, of course, identical. Even in this ternary schema, the differences between the 'meanings' behind real X occurrences are not discrete but continuous: senses can be more or less similar to each other, and this may or may not be reflected in lexicographic sources (dictionaries). One can follow this logic and define another option located between polysemy and identity: *context variance*, to use a term borrowed from Schlechtweg, Schulte im Walde, et al. (2018). In it, X_0 and X_1 still share the same sense, according to a dictionary, but are used in two significantly different contexts, making their perception by language users (and their typical associations) very different. This

[2] Or vice versa: the causality direction here is conceptually problematic and resembles the chicken or the egg dilemma.

[3] Note that we are not interested in cases when the word itself comes out of usage, even if its old sense is still active in the language (but now served by a different word); see the discussion of the distinction between semasiological and onomasiological changes in Chapter 3.

quaternary continuum in full can be described as follows (with the examples of X_0 and X_1 for English):

1. Homonymy:
    - ‘His *bark* was worse than his bite’
    - ‘He scratched the *bark* of the oak’
2. Polysemy:
    - ‘She submitted her *paper* to a journal’
    - ‘The report was printed on a piece of white *paper*’
3. Context variance:
    - ‘Careful *distancing* of blocks allow natural and controlled lighting for inner spaces’
    - ‘Self-quarantine and self-isolation are specific forms of social *distancing* in the period of the COVID-19 pandemic’
4. Identity:[4]
    - ‘The *crankshaft* rotates within the engine block through use of main bearings’
    - ‘Casting is today mostly used for *crankshafts* in cheaper, lower performance engines’

It is easy to distinguish between the two opposite extreme points of this continuum (identity and homonymy), but everything in between is continuous. Indeed, one can argue that the difference between identity (what is meant is exactly the same) and context variance (what is meant is the same in term of senses, but the context makes the word usage very different) is ill-defined, since it is difficult to come up with a precise test for this distinction. However, the same is true for the distinction between homonymy and polysemy: it is gradual as well (Kilgarriff, 1997). Dictionaries often make this decision based on whether X_0 and X_1 are related etymologically, but this factor is based more on historical factors than semantics. Overall, emergence of a new sense is a generative process, starting with word usages becoming more and more contextually varied, until at some (often rather arbitrary) point we decide that we observe a case of polysemy or even homonymy.

Thus, context variance is a span on the semantic proximity continuum and can be looked at as a semantic phenomenon. This is a long-standing position

[4] It should be noted that it is notoriously hard to come up with an example of an *absolutely* unambiguous word. One of (multiple) reasons for this is that the majority of words can be easily used in a metonymic or ironic sense, or simply metaphorically. Probably, some chemical substance term could be a better fit here, but we used a common name example, to avoid various complications surrounding the semantics of proper names.

in historical cognitive linguistics (Geeraerts, 1997). Consider, for example, the following quotation from Warren (1999):p. 219:

> ‘We must distinguish between the meaning of a word out of context and its meaning in context... The former type of meaning I refer to as dictionary meaning. It is the meaning that the lexicographer would be interested in. The latter type of meaning I refer to as contextual meaning which is the value we give a word in context. ...contextual meaning is part of *parole*’

Even if contextual meaning is indeed a part of *parole* in terms of Saussure (1916) or ‘linguistic performance’ in terms of Chomsky (1965) (while a dictionary meaning is part of *langue*), this ‘value we give a word in context’ is still meaning. Following this, we hold that systematic changes in word contexts (context variance) over time fall under the umbrella of ‘semantic shifts’ even if they are not accompanied with more or less discrete changes in the lexicographic senses. Some features of the meaning of a word can change enough to cause a significant drift in the words’ contexts but not enough to reach the point of acquiring a new ‘sense’.

In fact, if this were not the case, the second of the main lexical semantic change sub-tasks mentioned above (‘estimating and quantifying the degree of semantic change’) would be impossible to solve, since this degree can be inferred only from contextual cues in a corpus (including frequency of word usage in different contexts). This is especially true for tasks involving diachronic analogical reasoning and semantic relations, where the notion of ‘senses’ becomes even more blurred. However, we acknowledge that ‘sense-related’ understanding of semantic shifts is more established in the academic community, so we stick to calling them ‘semantic shifts proper’. They roughly correspond to the term ‘diachronic conceptual change’ from Tahmasebi, Borin, and Jatowt (2018). We still sometimes use the general terms ‘semantic shift’ or ‘semantic drift’ to denote the broader phenomena described above. The term ‘diachronic semantic change’ in this thesis covers all the mentioned cases with any lexical entities (and their relations) able to possess meaning.

1.2 Research questions

The present thesis systematically presents a set of research efforts which share the common topic of extracting semantic change data from word embeddings. Together, they demonstrate the diversity of use cases where the extracted data is employed to approach both theoretical linguistic problems and more practical natural language understanding tasks. Naturally, the thesis addresses several specific research questions. We present them below, along with abbreviated identifiers for easy referencing here and in the Conclusion.

The primary research question (*RQ0*) of this work asks **whether it is possible to reliably model diachronic semantic change using dense distributional word representations**. ‘To model’ here means ‘to capture important aspects of a phenomenon under scrutiny', which naturally includes the ability to automatically detect and/or predict manifestations of the phenomenon in the real world.

The primary question mentions a type of computational representation of meaning (distributional word embeddings), a linguistic phenomenon (diachronic semantic change in human languages), and algorithms to model the phenomenon using the computational representations. We first discuss and define the phenomenon, and then study the ability of the representations to capture its aspects. We do this by analyzing and evaluating both **different ways of training word representations on time-specific corpora** ('diachronic word embeddings'), and **different algorithms of extracting semantic change data from word embeddings**. The latter do not necessarily rely on *diachronic* word embeddings *per se*: for example, contextualized embedding based algorithms of semantic change detection described in Chapter 6 can pre-train word representations on large time-agnostic corpora and then infer time-specific token representations on time-specific corpora of lesser size. In any case, the main focus of our research lies in developing and studying algorithms of such extraction (this is actually what we call 'modeling of semantic change'), making it more complex than simply training word embeddings on time-specific corpora.

Our response to the primary question stated above is a series of case studies exploring how distributional word embeddings capture diachronic semantic change in English and other languages. Specific modeling algorithms can be efficient for different types, aspects and variations of semantic change. Chapters 4 and 6 cover semantic shifts proper, often manifested in words losing or acquiring lexicographic senses, while in Chapter 5, we mostly deal with semantic change of a more subtle type manifested in context variance.

As the primary question is being answered, it breaks into several smaller research questions, which logically group into three main themes:

- **Semantic change: defining key notions and surveying the field** (*RQ1*)

  The scientific field of automatic modeling of semantic change is extremely diverse and fragmented, even if we limit ourselves to the research employing distributional methods. The widespread lack of common terminology and awareness about previous work motivates this theme of the thesis.

  1. What are the main axes along which one can structure the current research on this topic? (*RQ1.1*)
  2. What were the primary related discoveries in recent years? (*RQ1.2*)

- **Answering linguistically oriented questions** (*RQ2*)

  This theme investigates the employment of distributional embedding based techniques to test linguistic hypotheses.

   1. Do evaluative adjectives change over time faster than other types of adjectives? Does this tendency hold across languages, based on corpus evidence? (*RQ2.1*)

- **Embedding-based algorithms of semantic change detection** (*RQ3*)

  This theme addresses questions related to the methodology of applying word embeddings to analyze temporal semantic change: both in general and for practical tasks.

   1. Can one use external datasets designed for other purposes as a proxy to evaluate semantic change detection algorithms based on word embeddings? (*RQ3.1*)
   2. Do word embeddings capture information about diachronic changes in semantic *relations* between words? (*RQ3.2*)
   3. What new perspectives on diachronic semantic change detection are brought with the recent contextualized embedding architectures? (*RQ3.3*)
   4. Do contextualized embeddings outperform static embeddings in this task, as they do in many other natural language processing areas? (*RQ3.4*)

We return to the same set of questions in our Conclusion (Chapter 7), where we summarize our answers to them.

1.2.1 Concurrency

As already mentioned, at the time of submitting this thesis (2019–2020) one can already find dozens of academic papers which lead to a positive answer to our primary research question. But the overwhelming majority of them appeared in the last two or three years, concurrently with the work on the thesis and well after 2015, when this work began. See Figure 3.6 for an outline of the development of the field of semantic change detection over time, and Chapter 3 in general for the detailed account of this evolution and its relation to my own research.

Additionally, the format of a conference (or even a journal) paper in natural language processing does not allow for answering this question on a large scale: as a rule, such papers focus on a single task. This is where we see the place of this thesis: to systematically explore abilities and limits of diachronic word embedding-based methods, employing results from different tasks, languages and aspects of semantic change, and comparing them to consistent baselines. We do not claim to cover all flavors of distributional embeddings, but instead make an attempt to enumerate (and to evaluate, wherever possible) at least the most important ones.

1.3 Thesis outline

The thesis is structured as follows.

We start in Chapter 2 ('Distributional modeling of meaning') with the description of data-driven distributional approaches to semantics, and particularly word embedding models. It also describes the shortcomings of embedding architectures and briefly presents other possible approaches to model meaning computationally. This chapter establishes the terminology and methods used throughout the thesis.

Next, Chapter 3 ('Modeling diachronic semantic change: state of the field') outlines the notion of semantic change both from a linguistic and from a computational point of view. It continues with the overview of the current approaches to tracing diachronic change using distributional models. This chapter both positions our work in the broad academic context and proposes foundations for structuring the field itself by determining the axes along which the research on the topic can be meaningfully compared.

Once the background knowledge is presented, Chapter 4 ('Measuring diachronic evolution of evaluative adjectives') moves to the linguistic problem of telling whether one lexical class is significantly different from another. In particular, we use diachronic word embeddings to quantitatively assess whether evaluative adjectives are more prone to shifting their meaning over time than other types of adjectives. This research was motivated by multiple examples of (mostly English) adjectives which significantly changed their meaning diachronically ('*awful*', '*terrific*', etc.). Crucially, here we conduct the same set of experiments across English, Norwegian and Russian language data, with the result being the same: no evidence is observed for evaluative adjectives to shift faster or more consistently than other adjective types. In this chapter, we do not propose new models for semantic change estimation, but test the applicability of the existing ones to a concrete linguistically motivated problem.

To move further, one needs a gold standard dataset for evaluating detection of semantic change. Before moving on to the specially designed diachronic semantic change test sets in Chapter 6, Chapter 5 ('Semantic change and real-world events: armed conflict dynamics') discusses how such a gold standard can be produced from historical armed conflict datasets. This approach is a form of distant supervision (Fang and Cohn, 2016), in which alternative data sources (not exactly describing the data under analysis) are used to train or evaluate a machine learning system. The chapter first describes the field of peace research in general and manually annotated armed conflict datasets in particular. It mostly focuses on the Uppsala Conflict Data Program (UCDP). The chapter continues with using this inferred gold standard to evaluate approaches to detecting semantic change by testing how good they are in capturing the start and end points of armed conflicts in time. These approaches are all based on word vector changes in distributional models trained on temporally annotated news texts. Successfully modeling detection of conflict state change allows us to actually mine armed conflicts dynamics data from raw texts in an unsupervised way. Note that in this chapter, we deal with semantic change belonging to the context variance part of the semantic proximity continuum: the representations of country names are drifting away from or towards the representations of conflict-related 'anchor words'.

Further in this chapter, we show how the same datasets can be used to evaluate the ability of diachronic word embeddings to capture changes in typed semantic *relations* between words. For this, we use the temporally labeled 'location–armed group' relations (for example, '*Afghanistan*'–'*Taliban*' in 2017 or '*Iraq*'–'*Islamic state*' in 2015). This task, which we refer to as 'temporal analogical reasoning' and formulate as filling in one-to-X relations, is itself one of our contributions, along with novel models for its handling. Even more importantly, we show that distributional representations do preserve information about semantic relations (and their changes) in the diachronic setup.

The case studies in the aforementioned chapters use 'static' word embedding models, in which each word is mapped to exactly one vector representation. This is a well-known shortcoming: ambiguous words receive representations which mix all their senses, often resulting in sub-optimal vectors. This is why we turn to more recent contextualized embedding algorithms employing deep neural networks: ELMo (Peters, Neumann, Iyyer, et al., 2018) and BERT (Devlin et al., 2019). In them, word representation at inference time depends on the input context, and thus ambiguity is handled naturally.

Chapter 6, shows how one can use contextualized word embedding models to analyze and estimate diachronic semantic shifts in word senses. This time, we employ specially designed human-annotated diachronic semantic change test sets to evaluate the our approaches: namely, those from the SemEval-2020 Shared Task 1 (Schlechtweg, McGillivray, et al., 2020) and the GEMS (Gulordava and Baroni, 2011). We propose several novel methods of estimating semantic change degree, managing to outperform all solutions submitted to the aforementioned shared task in the evaluation phase.

Additionally, in this chapter we describe some of the issues arising when one uses contextualized embedding architectures for semantic change detection. We analyze and categorize typical cases when high semantic change score is predicted by a system, but it is does not manifest a 'proper' semantic shift, as expected by a historical linguist. Despite these issues, empirical results of the approaches based on contextualized embeddings do outperform those based on static embeddings, which is important for future research in the field.

Finally, in Chapter 7, we conclude and summarize the results of the thesis. In particular, we describe in detail the contributions of each chapter and revisit the research questions enumerated above. Future work directions are outlined, which partially overlap with overall challenges facing the field of distributional semantic change modeling. We also list all publicly available code, trained models and datasets produced in the course of work on the thesis.

1.4 Publications

As this thesis was progressing, parts of it were published as peer-reviewed papers. In most cases, the venues were the ACL-sponsored conferences (ACL, EMNLP, EACL, CONLL, COLING, *SEM, SemEval, NODALIDA, etc) and their collocated workshops. For the chapters partially based on published work, we provide a footnote referring the reader to the corresponding paper(s) at the beginning of each chapter.

Chapter 2

Distributional modeling of meaning

This introductory chapter describes data-driven distributional approaches to computational semantics, and particularly word embeddings which became widely used throughout natural language processing in the 2010s. A reader familiar with the fundamentals of distributional semantics and word embedding models can skip this part and move directly to Chapter 3. However, this chapter is still important in that it establishes the terminology that will be used throughout the whole thesis. We also hope it can be useful as a focused overview of the field.

2.1 The distributional hypothesis

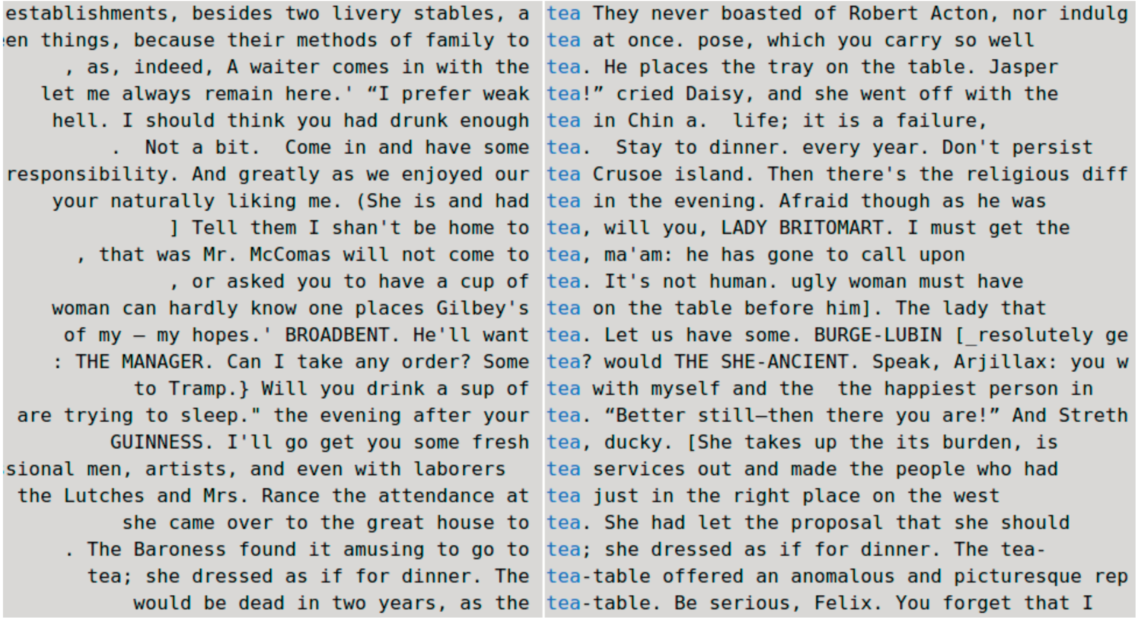
The distributional hypothesis is one of the most important notions in contemporary computational linguistics and natural language processing (NLP). It is the central theoretical foundation for distributional semantics: an empirical branch of linguistics concerned with lexical meaning. It is key in providing machine-readable word representations ('distributional models') for modern natural language understanding algorithms, mostly based on statistical data analysis and machine learning.

The distributional hypothesis was first formulated by Firth (1935); Harris (1954); Firth (1957), and by other linguists (sometimes it is claimed that a similar idea can be found in Ludwig Wittgenstein's texts from the 1930s). It is the idea that word meaning is characterized by its contexts. In a widely (over-)quoted statement, Firth (1957) puts it as '*You shall know a word by the company it keeps.*' Thus, words with similar typical contexts tend to have similar meaning.

The general concept becomes clear by looking at the Figures 2.1 and 2.2. One can see that the words nearby the words '*tea*' and '*coffee*' (their neighbors) in natural texts tend to be to some extent similar.[1] These are the words like '*some*', '*cup*', '*fresh*', etc. This illustrates that semantically similar words share similar contexts.

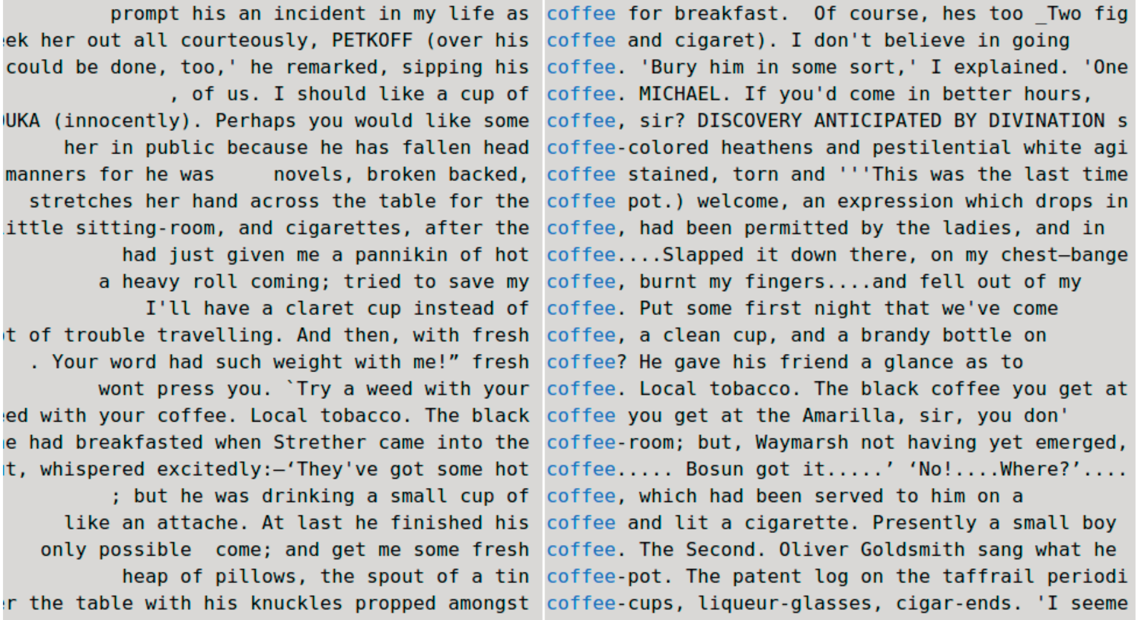
To be more exact, the distributional hypothesis postulates that co-occurrence statistics (word co-occurrence distributions) extracted from a large enough natural language corpus captures central aspects of the 'meaning' of words as perceived by humans. It is important that the meaning is inferred from textual data without using any external human knowledge, in a completely data-driven, unsupervised (or semi-supervised) way. The necessity to analyze

[1] Contexts were extracted from the *Project Gutenberg* English subcorpus (https://www.gutenberg.org/)



```
establishments, besides two livery stables, a tea They never boasted of Robert Acton, nor indulg
en things, because their methods of family to tea at once. pose, which you carry so well
, as, indeed, A waiter comes in with the tea. He places the tray on the table. Jasper
let me always remain here.' “I prefer weak tea!” cried Daisy, and she went off with the
hell. I should think you had drunk enough tea in Chin a.  life; it is a failure,
.  Not a bit.  Come in and have some tea.  Stay to dinner. every year. Don't persist
responsibility. And greatly as we enjoyed our tea Crusoe island. Then there's the religious diff
your naturally liking me. (She is and had tea in the evening. Afraid though as he was
] Tell them I shan't be home to tea, will you, LADY BRITOMART. I must get the
, that was Mr. McComas will not come to tea, ma'am: he has gone to call upon
, or asked you to have a cup of tea. It's not human. ugly woman must have
woman can hardly know one places Gilbey's tea on the table before him]. The lady that
of my — my hopes.' BROADBENT. He'll want tea. Let us have some. BURGE-LUBIN [_resolutely ge
: THE MANAGER. Can I take any order? Some tea? would THE SHE-ANCIENT. Speak, Arjillax: you w
to Tramp.} Will you drink a sup of tea with myself and the  the happiest person in
are trying to sleep." the evening after your tea. “Better still—then there you are!” And Streth
GUINNESS. I'll go get you some fresh tea, ducky. [She takes up the its burden, is
sional men, artists, and even with laborers tea services out and made the people who had
the Lutches and Mrs. Rance the attendance at tea just in the right place on the west
she came over to the great house to tea. She had let the proposal that she should
. The Baroness found it amusing to go to tea; she dressed as if for dinner. The tea-
tea; she dressed as if for dinner. The tea-table offered an anomalous and picturesque rep
would be dead in two years, as the tea-table. Be serious, Felix. You forget that I
```

Figure 2.1: Contexts of the word ‘*tea*’



```
prompt his an incident in my life as coffee for breakfast.  Of course, hes too _Two fig
ek her out all courteously, PETKOFF (over his coffee and cigaret). I don't believe in going
could be done, too,' he remarked, sipping his coffee. 'Bury him in some sort,' I explained. 'One
, of us. I should like a cup of coffee. MICHAEL. If you'd come in better hours,
UKA (innocently). Perhaps you would like some coffee, sir? DISCOVERY ANTICIPATED BY DIVINATION s
her in public because he has fallen head coffee-colored heathens and pestilential white agi
manners for he was     novels, broken backed, coffee stained, torn and '''This was the last time
stretches her hand across the table for the coffee pot.) welcome, an expression which drops in
ittle sitting-room, and cigarettes, after the coffee, had been permitted by the ladies, and in
had just given me a pannikin of hot coffee....Slapped it down there, on my chest—bange
a heavy roll coming; tried to save my coffee, burnt my fingers....and fell out of my
I'll have a claret cup instead of coffee. Put some first night that we've come
t of trouble travelling. And then, with fresh coffee, a clean cup, and a brandy bottle on
. Your word had such weight with me!” fresh coffee? He gave his friend a glance as to
wont press you. `Try a weed with your coffee. Local tobacco. The black coffee you get at
ed with your coffee. Local tobacco. The black coffee you get at the Amarilla, sir, you don'
e had breakfasted when Strether came into the coffee-room; but, Waymarsh not having yet emerged,
t, whispered excitedly:—‘They've got some hot coffee..... Bosun got it.....’ ‘No!....Where?’....
; but he was drinking a small cup of coffee, which had been served to him on a
like an attache. At last he finished his coffee and lit a cigarette. Presently a small boy
only possible  come; and get me some fresh coffee. The Second. Oliver Goldsmith sang what he
heap of pillows, the spout of a tin coffee-pot. The patent log on the taffrail periodi
r the table with his knuckles propped amongst coffee-cups, liqueur-glasses, cigar-ends. 'I seeme
```

Figure 2.2: Contexts of the word ‘*coffee*’

meaning with statistical methods was understood even 85 years ago: in the words of Firth (1935):p. 50, ‘in such subjects as semantics, which [. . .] is rather like meteorology, statistical and behaviouristic methods are widely held to be the only ones likely to take us further in our efforts to understand how language really works’.

Since the distributional meaning representations are inferred directly from texts, they naturally conflate *intensional* or ‘*denotational*’ meaning (manifested in dictionary definitions of ‘senses’ and traditionally considered to be the only part of meaning which is ‘really linguistic’) and *referential* or *non-denotational* meaning. The latter conveys information about particular members (referents) of the class covered by the denotational meaning, or some emotive and stylistic overtones (connotations). Non-denotational meaning is often also linked to ‘encyclopedic information’ or ‘world knowledge’. From a theoretical point of view, this conflation can be either beneficial or not, depending on the particular line of thought. In this thesis, we largely follow the opinion of Geeraerts (1997) that there is no clear borderline between the linguistic meaning and world knowledge. Conveniently, these words are written precisely in the context of diachronic semantic change (Geeraerts, 1997:p. 25):

> ‘...diachronic semantics has little use for a strict theoretical distinction between the level of senses and the level of encyclopaedic knowledge pertaining to the entities that fall within the referential range of such senses. In semantic change, the ‘encyclopaedic’ information is potentially just as important as the purely semantic ‘senses’ (to the extent, that is, that the distinction is to be maintained at all).’

From a practical point of view (solving real-world NLP tasks), the conflation of denotational meaning and encyclopedic information in distributional representations is almost always beneficial. The reason is that humans hardly can perceive or generate natural language utterances in any way which completely abstracts away from extra-linguistic ‘encyclopedic’ data or connotations and associations triggered by particular words. Processing limited to the denotational part of word meaning is possible only in very restrained artificial circumstances. Since most practical NLP tasks deal with attempts to approximate decisions made by human speakers, the conflation described above is again helpful. Still, it should be kept in mind when using distributional meaning representations. We discuss this issue in some additional details in section 2.5.

Formally speaking, within the distributional approach any linguistic entity can be represented as a vector of frequencies for this entity occurring together with other linguistic entities (its contexts) in a given corpus. In other words, lexical vectors are located in a semantic space with all the possible contexts or semantic features as dimensions (Osgood et al., 1964). The next section 2.2 discusses the details of such semantic spaces.

	vector	**meaning**	**hamster**	**corpus**	**weasel**	**animal**
vector	0	10	0	8	0	0
meaning	10	0	1	15	0	0
hamster	0	1	0	0	20	14
corpus	8	15	0	0	0	2
weasel	0	0	20	0	0	21
animal	0	0	14	2	21	0

Table 2.1: A simple example of a co-occurrence matrix.

2.2 Working with co-occurrence matrices

Together, the word vectors of a given word set (vocabulary) constitute a co-occurrence matrix, an example of which is shown in Table 2.1, with six words: ‘*vector*’, ‘*meaning*’, ‘*hamster*’, ‘*corpus*’, ‘*weasel*’ and ‘*animal*’ occurring near each other in an imaginary text. Words are represented with row vectors $\vec{\boldsymbol{x}} \in \mathbb{R}^6$: for example, ‘*hamster*’ is $[0, 1, 0, 0, 20, 14]$. Naturally, the word ‘*hamster*’ occurs in the vicinity of the word ‘*animal*’ (cf. phrases like ‘*hamster is an animal*’) much more often than in the vicinity of the word ‘*vector*’ (this paragraph is a notable exception). The same is true for ‘*weasel*’, giving us empirical grounds to state that the lexical meanings of ‘*hamster*’ and ‘*weasel*’ are close to each other.

In real-world distributional semantic models, the co-occurrences of each unique word in a corpus with all the other words within a given window[2] are counted. With the vocabulary V, each word a is represented with a high-dimensional vector $\vec{\boldsymbol{a}} \in \mathbb{R}^{|V|}$. The vector entries correspond to the other words of the corpus' vocabulary $(b, c, d...|V|)$. The values of the entries are frequencies of words co-occurrences $(a|b, a|c$, etc).

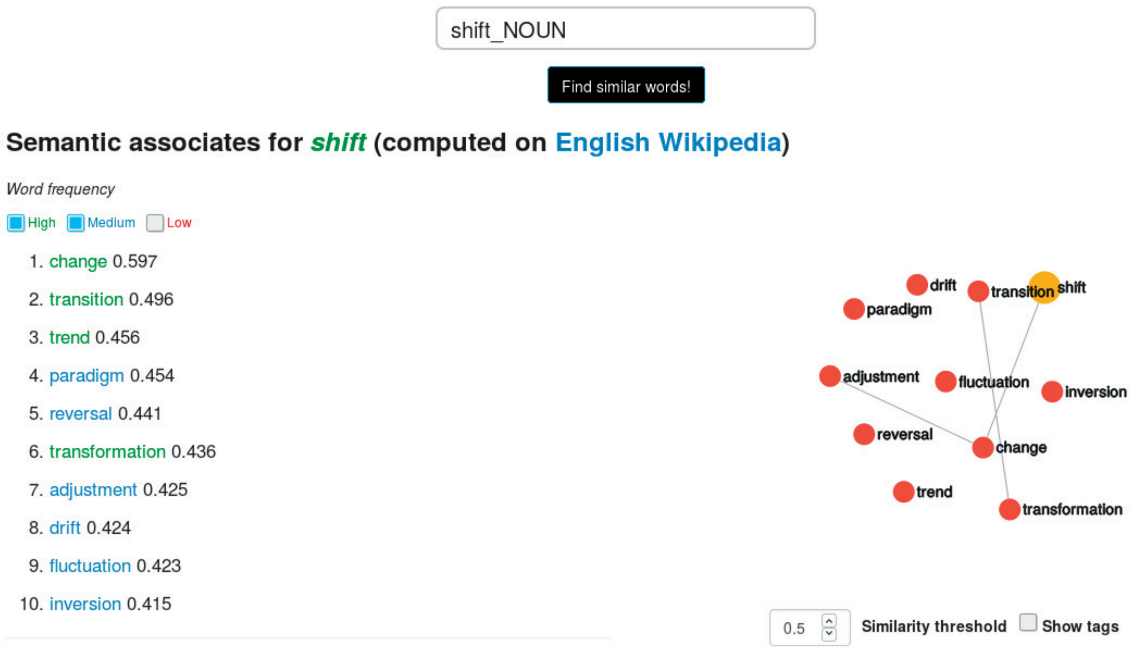
Semantically similar words tend to possess similar vectors (because they are used in similar contexts), while the non-related words' vectors are farther away from each other. Vector ‘similarity’ can be defined in many different ways: as Euclidean distance, dot product or cosine similarity, among others.

Cosine similarity is usually preferred as a measure of vector similarity. Conceptually, it is the cosine of the angle between two vectors (with the same origin) and takes values from -1 to 1. It lowers as the angle grows, and grows as the angle lessens. Formally it can be expressed as the dot product of unit-normalized vectors:

$$cos\,(\vec{\boldsymbol{w1}}, \vec{\boldsymbol{w2}}) = \frac{\vec{\boldsymbol{w1}} \cdot \vec{\boldsymbol{w2}}}{|\vec{\boldsymbol{w1}}||\vec{\boldsymbol{w2}}|} \tag{2.1}$$

When vectors point in the same direction, $cos = 1$, when they are orthogonal, $cos = 0$, and when they point at the opposite directions, $cos = -1$. It is important

[2] Ranging from ‘one word to the left and one word to the right’ to ‘the whole document’.



shift_NOUN

Find similar words!

Semantic associates for *shift* (computed on English Wikipedia)

Word frequency

☑ High ☑ Medium □ Low

1. change 0.597
2. transition 0.496
3. trend 0.456
4. paradigm 0.454
5. reversal 0.441
6. transformation 0.436
7. adjustment 0.425
8. drift 0.424
9. fluctuation 0.423
10. inversion 0.415

0.5 Similarity threshold □ Show tags

Figure 2.3: The nearest neighbors of the English noun '*shift*' in the distributional model trained on the English Wikipedia.

that cosine similarity inherently smooths the influence of vector magnitudes (thus avoiding assigning too high similarity scores to frequent words, which is the case when using the simple dot product). Often, all the vectors in the matrix are unit-normalized before any operations with them, since in this case cosine similarity boils down to calculating the dot product. The *cosine distance* measure is sometimes used, which is a simple inversion of cosine similarity ($1 - cos$).

Cosine similarity provides an efficient way of measuring semantic similarity between words. By ranking all the words in the vocabulary by the cosine distance of their vectors to the vector of the query word, one can easily find the query word's nearest neighbors: n words with the highest similarity to it. Linguistically, the nearest neighbors are the words paradigmatically related to the query, the ones by which the query word can be substituted in natural language utterances.

The list of the word's nearest neighbors can itself tell a lot about what does the word means, as shown in Figure 2.3, with the query word '*shift*'. This is a screen shot from our WebVectors web service (Fares et al., 2017)[3], featuring a distributional model trained on the English Wikipedia. It lists 10 words with the highest values of cosine similarity to the vector of '*shift*' (the corresponding cosine similarities are given near each neighbor). The right part of the screen shot displays the '*shift*' ego graph, where nodes are its nearest neighbors, connected with edges if their pairwise similarity exceeds the user-defined threshold (0.5 in

[3] http://vectors.nlpl.eu/explore/embeddings/

this case).

Because of the Zipfian (power law) distribution of word frequencies in natural languages (Zipf, 1949), some words can often be observed in the vicinity of other words simply because they are frequent, not because these co-occurrences are really indicative of any meaning. This is true, for example, for English determiners '*the*' and '*a*' which can co-occur with practically any noun.

In order to filter this noise, various weighting schemata are applied to absolute co-occurrence counts extracted from corpora. As a rule, they draw on variations of information theoretic association measures quantifying the probability of the $a|b$ co-occurrence being accidental, given the individual and joint frequencies of a and b. The most widespread association measure in NLP is arguably positive point-wise mutual information (PPMI) introduced by Church and Hanks (1989) and given in Equation 2.2.

$$PPMI(a,b) = \max\left(\log_2\left(\frac{P(a,b)}{P(a)P(b)}\right), 0\right) \tag{2.2}$$

Arguably, the first 'traditional' distributional vector space model put to practical use was the Term Frequency - Inverted Document Frequency (tf/idf) statistic introduced by Jones (1972). Since then, such models have been developed and studied for decades, and have become widely utilized in many natural language processing applications, from simply measuring lexical semantic similarity to inferring complex topical structures of document collections. It is impossible to survey all the variations in this thesis, and we refer interested readers to the extensive review by Turney, Pantel, et al. (2010). To name only a few relatively recent ones:

- *Latent Semantic Analysis* (LSA) or *Latent Semantic Indexing* (LSI) (Landauer and Dumais, 1997) was particularly successful in information retrieval, where it powered search for similar documents;
- Later, LSA was extended to *Probabilistic LSA* (PLSA), based on a version of the expectation-maximization (EM) algorithm (Hofmann, 1999);
- More distantly related is the *Latent Dirichlet Allocation* (LDA) model from Blei, Ng, et al. (2003), which allows one to cluster words and documents by their topics;
- The *Random Indexing* algorithm (Kanerva et al., 2000; Sahlgren, 2005; Velldal, 2011) employs an incremental formulation of random projections and is directly related to the dimensionality reduction problem described in Section 2.3.

Some of these approaches are document-centric (LSA and LDA), while others are word-centric (Random Indexing), which means they use different types of contexts. Still, all of them are based on the idea of employing distributional signal from raw texts to capture information about the meaning of linguistic entities (either documents or words). In this thesis, we deal with representations for words and with the possibility to use them for the modeling of semantic change.

2.3 The curse of dimensionality

The most widespread type of contexts in distributional models is word neighbors, which means that the set of all possible contexts generally equals the size of the vocabulary of the corpus, and it can be quite large (sometimes in the order of millions). Additionally, the vectors tend to be *sparse*, with most entries being zeros, because of the Zipfian word frequencies distribution in all natural languages.

Sparse representations are usually inefficient: for example, even a 50-words text will have to be represented by a vector $\vec{\boldsymbol{x}} \in \mathbb{R}^{100000}$, if there are 100 000 words in the vocabulary. There exist ways of efficiently handling sparse matrices; however, from the very beginning of vector space meaning representations, researchers tried various ways of turning them into *dense* vectors with dimensionality much lower than the vocabulary size – compensated by the fact that there are no zero entries (Bullinaria and J. P. Levy, 2007). Such dense vectors are commonly called 'embeddings', reflecting the idea that we attempt to 'embed' or 'project' high-dimensional entities into a low-dimensional space.

One common way to achieve this aim is to first build a standard co-occurrence matrix like the one in Table 2.1, and then find a low-rank matrix that approximates the original one best. This approach is called 'dimensionality reduction' and there exists a wide variety of well-developed techniques for that: for example, *Principal Components Analysis* (PCA) by Pearson (1901), *Singular Value Decomposition* (SVD) by Golub and Reinsch (1970), or *Locality Sensitive Hashing* (LSH) by Van Durme and Lall (2010). One can approximate a high-dimensional matrix $\boldsymbol{A} \in \mathbb{R}^{x}$ with a low-dimensional matrix $\boldsymbol{B} \in \mathbb{R}^{y}$ (where $y \ll x$), while still preserving more or less the same similarity relations between the vectors. The downside is that the vector entries or components are not interpretable any more, like they were before the reduction. In other words, after the reduction, they do not correspond to any particular linguistic entities.

Reducing the dimensionality of semantic vectors is also important from the point of view of visualization. Most humans can't easily imagine the relations between vectors of dimensionalities higher than three. Thus, it is crucial to embed high-dimensional vectors into two- or three-dimensional projections, which humans can physically inspect. Figure 2.4 shows a 2-dimensional embedding which was created using the t-SNE algorithm (Van der Maaten and Hinton, 2008). The plot axes represent the components of this embedding. The original 300-dimensional vectors for the 10 words in the plot were extracted from the distributional model trained on the Google News corpus (Mikolov, Sutskever, et al., 2013). Even with the dimensionality reduced to two, it is easy to see how the model captures the semantic similarity between the words '*town*', '*city*' and '*capital*'. Additionally, it allows one to get a notion of 'semantic directions' within distributional models: the imaginary lines between the names of the countries

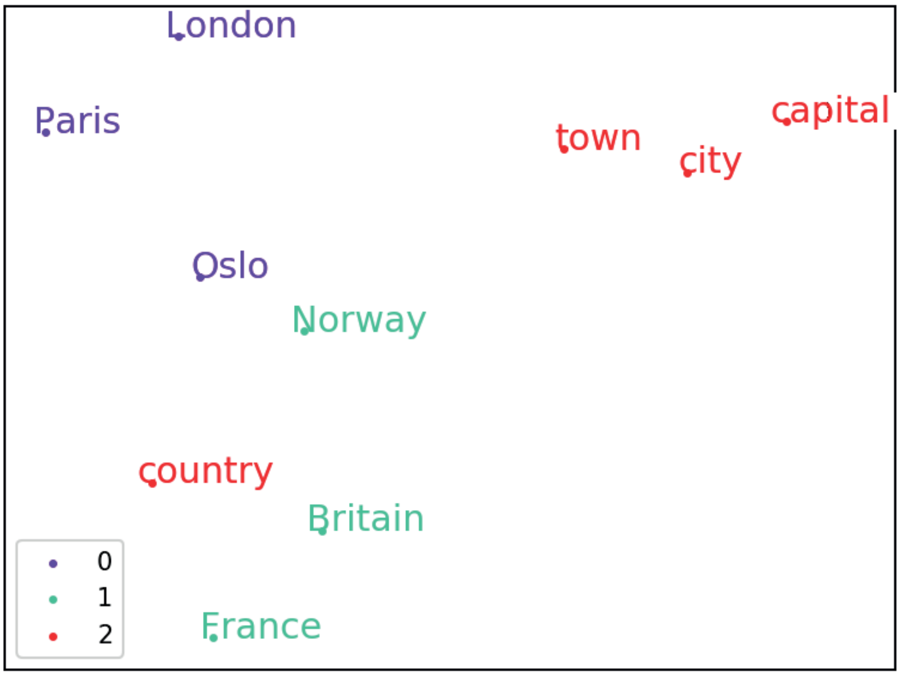


Figure 2.4: Two-dimensional t-SNE projection of 10 word vectors from the distributional model trained on the Google News corpus. The colors were added manually to easily distinguish the lexical groups containing 0) capitals, 1) countries, 2) common nouns.

and the corresponding capitals (‘*France*’ to ‘*Paris*’, ‘*Britain*’ to ‘*London*’ and ‘*Norway*’ to ‘*Oslo*’) would be almost parallel. We will use this important property of distributional models in Chapter 5 of this thesis.

More recently, another type of distributional models appeared, using machine learning techniques to approach the dimensionality reduction problem from a different angle. This is the kind of algorithms we primarily use in this thesis, and we survey them in the next section.

2.4 Rise of machine-learned distributional models

After 2013, distributional semantics enjoyed substantially growing attention because of the emergence of a particular class of approaches, which Baroni et al. (2014) dubbed ‘prediction-based models’ (since the vectors here are often optimized for *predicting* neighboring words) and opposed to the ‘traditional’ ‘count-based models’ described in Section 2.2.

The ‘predict vs. count’ dichotomy itself is not flawless, since in fact some algorithms show properties belonging to both approaches, like, for instance, *GloVe* from Pennington et al. (2014), in which a full co-occurrence matrix is first built, but then a prediction-like method is used to create the actual dense vectors. Another example is *Random Indexing* from Kanerva et al. (2000), which seems to belong to neither type: in this case, at no point a full co-occurrence matrix is constructed, but at the same time there are no 'predictions' of any kind. More meaningful distinctions can be made: for example, the one between *explicit representations*, where vectors are high-dimensional and sparse, but directly interpretable (corresponding to contextual features one-to-one) and *continuous representations* (embeddings), where the dimensionality is reduced and the resulting dense representations are not interpretable any more. However, the term 'prediction-based models' has already become quite ubiquitous, especially when talking about the *word2vec* algorithms (see below). They exhibit a promising set of properties and yield state-of-the-art performance in many NLP tasks.

The general idea of the prediction-based algorithms is that they *approximate* the co-occurrence data instead of calculating them directly. Word vectors are *trained* on the textual data using machine learning approaches, with the objective to maximize the similarity between the paradigmatic neighbors found in the corpus, while minimizing the similarity for unseen contexts. In a sense, this is a special case of training a language model which predicts the next word in a sequence, and word embeddings are created as a by-product of this training, as first shown by Bengio et al. (2003). Thus, word co-occurrences in the training corpora are used as a (weak) supervision signal, but the whole co-occurrence matrix is never actually constructed (although there are exceptions, like the already mentioned *GloVe*).

Vectors are first initialized randomly and then gradually converge to the optimal values at the training time, as we move through the corpus with a sliding window of a predefined length and try to predict neighbors on the basis of the current word. As Rong (2014) puts it, 'the vector of a word *w* is 'dragged' back-and-forth by the vectors of *w*'s co-occurring words, as if there are physical strings between *w* and its neighbors [. . .] like gravity, or force-directed graph layout.' In other words, during the training, each time the model sees a word in context, it makes a guess and shifts this word's vector a bit closer to the vectors of its neighbors, based on the resulting prediction error (loss). After millions and billions of such small updates, vectors converge to the state which best reflects the semantic similarities between words in the training corpus.

Prediction models often employ artificial neural networks (Goldberg, 2017). In particular, Mikolov, Sutskever, et al. (2013) introduced the highly efficient Continuous Skip-gram and Continuous Bag-of-Words (CBOW) algorithms. Essentially, they proposed a modification of the already existing feed-forward neural networks language modeling techniques and found the most efficient combination of hyperparameters. At training time, CBOW learns to predict the current word based on its context, while Skip-gram learns to predict context based on the current word. The differences between two architectures are shown in Figure 2.5. At each training instance, the input for the prediction is:

- CBOW: average input vector for all context words. We check whether the current word output vector is the closest to it among all vocabulary words.

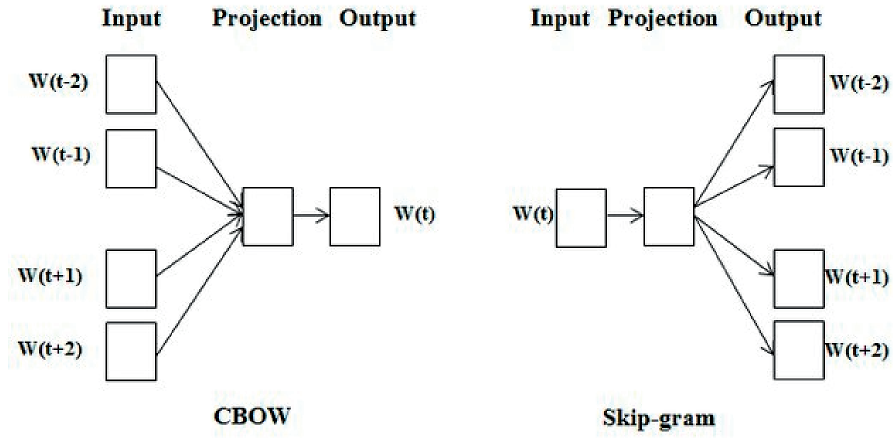


Figure 2.5: The now-famous depiction of the two *word2vec* architectures (CBOW ans Skip-gram) from Mikolov, K. Chen, et al. (2013).

- Skip-gram: current word input vector. We check whether each of context words output vectors is the closest to it among all vocabulary words. Sometimes, the SGNS abbreviation is used to refer to this algorithm, meaning 'SkipGram with Negative Sampling'.

This 'closeness' is calculated with the help of the dot product (or cosine similarity) and then turned into probabilities using softmax. During the training, the model updates two weight matrices: of input vectors ($\boldsymbol{W}^I$, from the input layer to the hidden layer) and of output vectors ($\boldsymbol{W}^O$, from the hidden layer to the output layer). As a rule, they share the same vocabulary, and only the input vectors are then used at test time as a look-up table for word embeddings in practical tasks.

The network architecture in both algorithms is very shallow, with a single hidden/projection layer between the input and the output layers. The training objective is to maximize the probability of observing the correct output word(s) w_t given j context word(s) $c_1 ... c_j$, with regard to their current embeddings (sets of weights in the matrices). The loss function L is cross-entropy. For CBOW it is formulated as in Equation 2.3, and for Skipgram as in Equation 2.4. The learning itself is implemented with stochastic gradient descent.

$$L = -\log\left(P\left(w_t | \sum_{i=1}^{j} c_i\right)\right) \tag{2.3}$$

$$L = -\sum_{i=1}^{j} \log\left(P\left(c_i | w_t\right)\right) \tag{2.4}$$

O. Levy and Goldberg (2014) and O. Levy, Goldberg, and Dagan (2015) showed that Skip-gram implicitly factorizes a word-context matrix of PPMI coefficients. Implicitness is important here: since the matrix is never explicitly constructed, the algorithm is more computationally efficient than the previous methods, especially with regards to memory requirements. However, the resulting vectors still approximate the rows of the same word-context matrix, even if it never materializes in memory as is. This further supports the claim that the differences between ‘count-based’ and ‘prediction-based’ distributional models are not as well-founded as it was customary to think when dense word embeddings first appeared.

Regardless, fast and high-quality training of word embeddings on huge amounts of texts was made possible after Mikolov, Sutskever, et al. (2013) released a ready-to-use *word2vec*[4] tool with its source code, which ensured reproducibility and ease of use. Various improvements and implementations in several programming languages quickly followed. Nowadays, these algorithms can be found in libraries like Gensim[5] (Řehůřek, 2011), TensorFlow[6] (Martin Abadi et al., 2015), Keras[7] (Chollet et al., 2015) and PyTorch[8] (Paszke et al., 2019), among many others. Still, these approaches (and the distributional hypothesis that they are based on) have several serious shortcomings. We discuss them in the next Section 2.5.

2.5 Shortcomings of word embeddings

This section presents important limitations which should be kept in mind when working with word embeddings:

1. Technical issues (data and compute requirements, etc.).
2. Interpretability issues.
3. The need to align different embedding spaces.
4. Conceptual problems with the distributional hypothesis itself.
5. The ability of word embeddings to represent word senses.

Some of the shortcomings of word embedding-based approaches to represent meaning are purely technical. Among others one can mention that they require training corpora of a very large size: small text collections do not provide sufficient variation for the distributional representations to become sufficiently general. However, this is not something specific to word embeddings: all modern machine learning algorithms are data-hungry. Also, obtaining relevant raw text corpora is still as a rule much easier and cheaper than the manual labor implied by the knowledge-based approaches (dictionaries, ontologies, etc). The same can be said about computational power requirements. It is true that distributional

[4] https://code.google.com/archive/p/word2vec/
[5] https://github.com/RaRe-Technologies/gensim
[6] https://www.tensorflow.org/
[7] https://keras.io/
[8] https://pytorch.org/

representations are compute-intensive, but the methods of training them are now well-developed and robust; as a rule, creating or obtaining useful word embeddings is not much of a problem. *Word2vec* vector representations can be trained on a billion-word size corpus in a matter of hours using a standard laptop. However, with the recently introduced contextualized embeddings (see Section 2.6), this issue becomes important again: such models can be slow and expensive to train even when using GPU-accelerated computation.

The technical shortcomings mentioned above are true for all word embeddings: both explicit and continuous, either created from a real word co-occurrence matrix or 'trained' using a prediction-based approach. But there are issues manifesting themselves only for specific embedding types. A notorious example is the problem of interpretability relevant for continuous word vectors. Unlike the explicit representations (where each vector component corresponds to a known context word), continuous or dense vectors are black-boxes (Linzen et al., 2019): it is difficult to match their components to any meaningful linguistic feature. Essentially, this is a mismatch between the continuous nature of prediction-based word embeddings (where information is distributed across many vector components) and the discrete nature of language. In response to this, it can be argued that discreteness is rather a property of traditional language descriptions than of the language itself, but it will be difficult to support either point of view with empirical evidence. Anyway, the lack of interpretability can certainly be a problem if some explanation for the system predictions is required.

Another issue concerns the 'trained' prediction-based word embeddings. It is related to their stochastic nature, and we discuss it in the next subsection.

2.5.1 Aligning stochastic representations

Training prediction-based word embedding models on text corpora is rather straightforward. However, it is not that straightforward to compare vectors for one and the same word across different models, which is required if one is analyzing the differences between embeddings trained on different corpora. This includes time-specific corpora, which directly concerns our topic of diachronic semantic change.

It usually makes no sense to, for example, directly calculate cosine similarities between embeddings of one and the same word in two different sets of embeddings. It is true even if the embeddings are estimated by the same underlying algorithm using the same hyperparameters. But the problem will be even more severe when it comes to comparing embeddings produced by different algorithms or with different hyperparameters. The reason is that most modern word embedding algorithms are inherently stochastic: the resulting vector sets are heavily dependent not only on the training data itself, but also on the original randomly initialized vector components (weights). Moreover, random choice of negative examples (in *word2vec* and its derivatives) and the order of training data instances (when shuffling is used) add up to the non-deterministic nature of prediction-based word embeddings.

Thus, even when trained on the same data, different runs will produce slightly different weights in the models (though with roughly the same pairwise similarities between word vectors, since the models will be invariant under rotation). This is even more expressed for models trained on different corpora. It means that even if word meaning is completely identical in two training corpora, the direct cosine similarity between its vectors trained on these two corpora can still be quite low, simply because the random initializations of the two models were different.

This is not so much of an issue for explicit and count-based algorithms, since they are deterministic and do not use any randomness while creating word representations. Given that one and the same vocabulary of context words is used, two count-based models trained on two different corpora will be fully comparable (even after applying some dimensionality reduction algorithm like SVD, etc). But if one would like to compare vectors from two prediction-based embedding models, she will have to first employ some *alignment* technique.

The need for alignment arises not only in the context of semantic change modeling. The same task is being handled, for example, in the field of cross-lingual embeddings (Ruder et al., 2019), where one has to map vectors of words in the language A to the vectors of their translation equivalents in the language B. More generally, alignment is inducing a shared semantic space from several different spaces. This problem can be solved in a variety of supervised and unsupervised ways, with the researchers in cross-lingual embeddings preferring supervised approaches, where a small bilingual dictionary is used as a seed to induce the mapping they need. In the field of monolingual semantic change detection, the problem is less difficult, since the intersection of the vocabularies of the vector spaces under analysis is usually quite large, and unsupervised methods like Orthogonal Procrustes (Gower, Dijksterhuis, et al., 2004) can be used. We describe various ways to align word embedding models (or to make them comparable via other means) in the next chapter in Section 3.2.5, specifically in the context of diachronic semantic change modeling.

2.5.2 Scientific credibility of the distributional hypothesis

Apart from technical ones, there are several *conceptual* issues with the distributional hypothesis. In this subsection, we focus on one of them. This issue is related to a simple (but tricky) question: **Is the distributional hypothesis really a hypothesis?** That is, can it be properly falsified? We argue that it is better to understand it more as a *useful assumption*.

Indeed, there are many examples of the distributional hypothesis working. By 'working' we mean that it was repeatedly and rigorously shown that technologies based on the distributional hypothesis tackle problems more efficiently than other approaches. So, there are many confirmations for it to be true, starting at least from Rubenstein and Goodenough (1965). In many publications it is silently implied that distributional representations literally manifest meaning.

However, this does not mean that the distributional hypothesis is *falsifiable* in Karl Popper's sense (Popper, 1962). To conform to this definition, a hypothesis

word1	**word2**	**output**	**gold**	**diff**
‘*girl*’	‘*maid*’	7.72	2.93	-4.79
‘*happiness*’	‘*luck*’	6.59	2.38	-4.21
‘*crazy*’	‘*sick*’	7.49	3.57	-3.92
‘*arm*’	‘*leg*’	6.74	2.88	-3.86
‘*breakfast*’	‘*supper*’	8.01	4.40	-3.61

Table 2.2: Example of the SimLex999 semantic similarity test set from Hill et al. (2015), along with predictions from a distributional model (the ‘output’ column).

should allow some way to prove that it is wrong. ‘To prove wrong’ here means ‘to find that the theory contradicts the known facts’. In our case, it would mean finding some way to reject the hypothesis that meaning is context. Is this possible at all? The answer to this question is far from being clear, considering the fact that the distributional hypothesis is entirely inductive: that is, based on observations, not on logical conjectures.

The root of the problem lies in the obvious fact that unfortunately we cannot observe meaning directly and empirically. Even with the help of brain imaging techniques, we still can’t reliably extract from the human mind, for example, the ‘meaning’ of the word ‘*dog*’ (Søgaard, 2016; Auguste et al., 2017). Any estimates of the meaning of words are necessarily only proxies. In fact, when one wants to intrinsically evaluate two sets of distributional representations (produced from different corpora or with different hyperparameters) and find out which is better, one has to rely on such a proxy, for example, on semantic similarity test sets. This method is often criticized (Faruqui, Tsvetkov, et al., 2016; Chiu et al., 2016), but still remains the most widely used.

Semantic similarity datasets contain human judgments about the semantic similarity of words (Hill et al., 2015). They usually come in the form of word pairs with some score of semantic similarity for each (for example, from 0 to 10, where 0 denotes completely unrelated words, and 10 denotes full synonyms). Each pair is scored by several informants, and their scores are averaged, to ensure reliability and robustness of scores.

Given such a dataset and a distributional model (a set of word representations), we can experimentally evaluate the model against the dataset. The evaluation is extremely simple and consist of producing the model’s predictions on the similarity of word pairs in the dataset (for the vector-based models, it boils down to calculating cosine similarity between word vectors). These estimates are then compared to those in the dataset (gold scores), and the Spearman rank correlation coefficient is calculated. The correlation close to 1 suggests that the model reproduces human judgments almost perfectly, while the correlation close to 0 would imply that the model’s predictions are close to being random.

For example, Table 2.2 shows the models' predictions in the 'output' column, and the human judgments in the 'gold' column. As we can see, in this case, this particular model tends to overestimate semantic similarities between words. It also predicts that the '*girl, maid*' pair is more similar than the '*crazy, sick*' pair, contrary to human judgments, etc.

One might think that such evaluation methods provide us with a way to falsify the distributional hypothesis. Unfortunately, they do not. There are several reasons for that.

The first reason is that any correlation score (not equal to 1 or 0) can be looked at either positively or negatively, depending on the subjective opinion of the observer. What level of correlation will command us to reject the distributional hypothesis altogether and why? Figure 2.6 presents the results of an experiment on the RuSimLex965 semantic similarity dataset for Russian (Kutuzov and Kunilovskaya, 2017), with dots representing word pairs. On the horizontal axis are the models' predictions about semantic similarities, and on the vertical axis are the human judgments (absolute values are converted to ranks). There is no absolutely clear trend, and in fact the Spearman ρ in this case is only about 0.4.

However, the word embeddings in question[9] perform well for many practical NLP tasks. They are of course not absolutely perfect, but this is not necessary for production usage. And even from the academic point of view, we can say that distributional models are inherently stochastic, statistical and even if they make mistakes in some cases, is does not disprove the theory in general.

The second (and even more important) reason is that there is no single semantic similarity (or any other intrinsic evaluation) test set being good for all. Such test sets are created by independent research groups, each with its own guidelines and principles of word selection. It is clear that the performance of distributional representations would critically depend on a tremendous amount of various properties of the test set, and the model performing poorly on one set can be superior on another. It is extremely difficult to find out what test sets are most representative of this or that language in general (Bakarov et al., 2018).

It is impossible to falsify the distributional hypothesis itself by evaluating distributional models against human judgments datasets. The distributional hypothesis does not forbid particular models to be 'good' or 'bad' on particular test sets, so it seems invincible from that side.

And of course, another difficult question is whether it is 'meaning' at all, that is conveyed by distributional patterns. Boleda and Erk (2015) say that '...a more general characterization of what distributional inference is and what purposes it can serve remains to be done'; Bender and Koller (2020) explicitly claim that real meaning cannot be learned from form alone (and this is exactly what distributional algorithms are trying to do). As if replying to this, Sahlgren (2008) proposes that semantics should be looked at as simple interplay of syntagmatic and paradigmatic relations between words, similar to the structuralist point of view. In this light, the distributional hypothesis at least receives something

[9] It was the Russian `ruscorpora_upos_skipgram_300_10_2017` model from our RusVectōrēs project (https://rusvectores.org/en/models/).

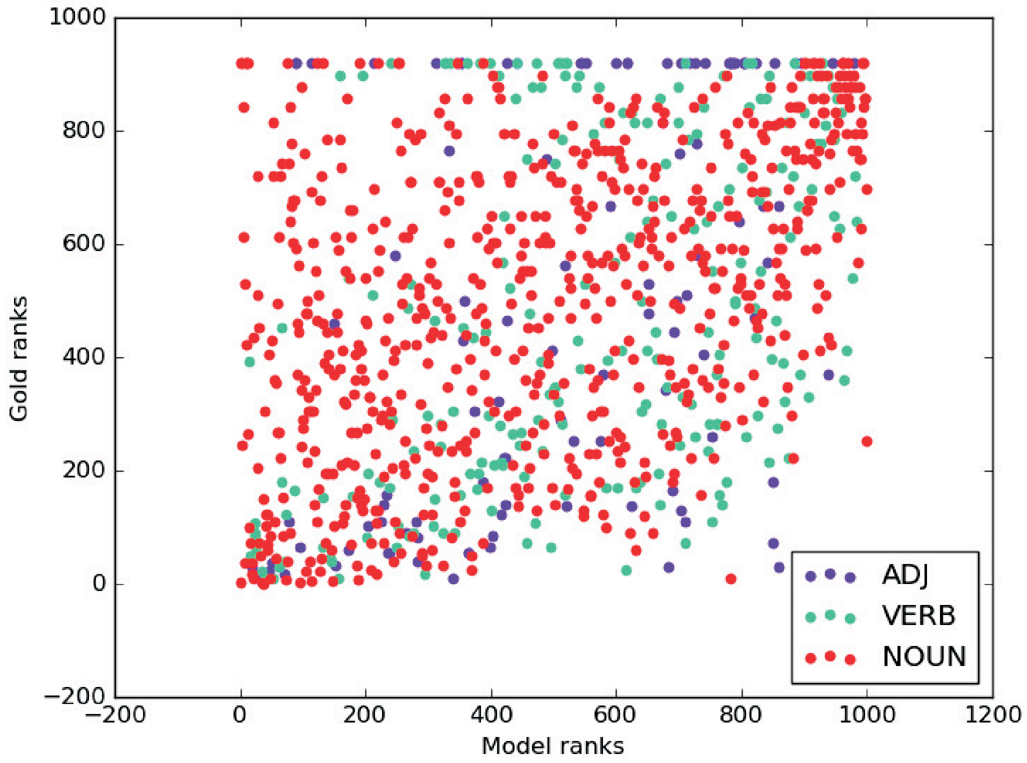


Figure 2.6: Interplay between gold human judgments and distributional model predictions on lexical semantic similarities.

resembling a solid theoretical foundation. However, even Sahlgren (2008) does not deal with the issue of falsifiability.

The problem seems to be related to the complex nature of the 'meaning' notion itself. The distributional hypothesis is to some extent recursive: when one says that 'meaning is distribution', one in fact says that 'distribution is distribution'. Due to this circularity, it is to some extent misleading to call it a 'hypothesis'. One indeed can suggest and test local hypotheses of particular distributional representations being better than others on particular test sets. But the global 'distributional hypothesis' itself is probably better be called a very useful *assumption*, which is extremely helpful in solving practical problems (one simply takes this assumption for granted). This statement might look obvious. However, as stated above, in many cases researchers tend to overlook the innately non-direct character of distributional meaning representations. We believe it is important to keep it in mind when working with such approaches.

2.5.3 Can word embeddings represent senses?

Another important issue with distributional meaning representations is handling word senses. Recall that a lexicographic 'sense' is essentially a word entry in a dictionary, where one and the same word form can have several entries, corresponding to different 'senses' of this word and together forming its 'conventional meaning'.[10] Can a continuous embedding for the word Z tell us how many senses does Z have and what are these senses?

The simplest answer to this is negative: it cannot. As already mentioned in the introduction, the whole notion of lexicographic word senses implies *discreteness*, and our Z embedding is *continuous*. Besides this conceptual barrier, there is a purely technological obstacle: even ambiguous words receive only one distributional embedding, which conflates all their senses into one. This results in distributional models being unable to capture polysemy and colexification in general (Yaghoobzadeh and Schütze, 2016).

Thus, simple attempts to infer discrete human-defined senses from a continuous vector trained on raw text will always be ad-hoc and incomplete (unlike knowledge-based methods which rely on dictionaries and ontologies). However, the NLP community keeps attempting, and there is some limited success in the task of modeling senses with word embeddings. One can mention the methods which involve clustering of averaged vectors for context words, thus creating *sense embeddings*, which can further be used to detect what sense the word is used in in a particular utterance (Schütze, 1998). Another family of approaches induces a sense inventory from pre-trained word embeddings via clustering of ego-networks of their nearest neighbors (Pelevina et al., 2016; Logacheva et al., 2020). Finally, multiple attempts were made to combine topic modeling and other non-parametric architectures with word embeddings for sense representation; see (P. Liu et al., 2015; Bartunov et al., 2016), among many others.

But the most important step in this direction was arguably the introduction of *contextualized embeddings* (Melamud et al., 2016); see also Section 2.6. They are not even really embeddings (in the sense of a simple vector lookup table): instead, they are full-fledged neural language models taking word sequence as an input and producing context-dependent word vectors. These 'contextualized' vectors are supposed to reflect word senses: '*mouse*' in the sentence '*I ordered a mouse for my laptop from Amazon*' will receive a different vector from '*mouse*' in the sentence '*When the mouse laughs at the cat, there is a hole nearby*'. Although such architectures are fully unsupervised and do not use any external knowledge about word senses, they achieve very competitive performance in word sense related tasks like word sense disambiguation and word sense induction (Pilehvar and Camacho-Collados, 2019; Loureiro and Jorge, 2019). In Chapter 6 we will show that contextualized embeddings indeed capture information related to word senses as defined in manually built ontologies, and apply them to lexical semantic change modeling. We also will present important issues which arise

[10] Note that other, more data-driven definitions of 'sense' are possible, for example, that word senses are 'abstractions over clusters of word usages' (Kilgarriff, 1997).

when one attempts to trace word sense changes with contextualized embeddings. We argue that the 'senses' they capture are more similar to 'readings' in the prototype theory as described by Geeraerts (1997), but this is not necessarily a bad thing.

Overall, the shortcomings of word embeddings described in this section do not prevent us from using the distributional approach throughout this thesis, but still should be kept in mind.

2.6 Recent trends in distributional semantic modeling

Nowadays, it is difficult to imagine any large-scale application dealing with human language (either in research or in industry) which does not use word embeddings in at least some parts (Desagulier, 2017). Google Translate employs word and sentence embeddings to generalize better when it analyzes the meaning of the source text (M. Johnson et al., 2017); Facebook employs word embeddings when it assesses semantic similarity of two posts, in order to decide which one to show, where and when (Bojanowski et al., 2017; J. Johnson et al., 2019).

Digital texts today are cheap to obtain and process, and distributional models now are trained on text collections containing billions of words in them (e.g., the whole Wikipedia, large news collections or simply millions of pages crawled from the Web). Although more textual data does not necessarily mean a better model, it almost always means a more diverse model with a better coverage.

Word vectors are used as input to complex artificial neural networks, greatly increasing their performance and widely replacing discrete word identifiers as input features (Goldberg, 2017). Traditional word representations were high-dimensional, sparse and categorical. Nowadays word embeddings are used almost exclusively, being comparatively low-dimensional, dense, continuous and distributed. Sometimes, such dense representations are learned simultaneously with training a larger neural network for a particular natural language processing task (sentiment analysis, machine translation, etc), as a dynamic part of such a network. However, another popular approach is to use pre-trained word representations provided by a third party (usually trained on a very large text corpus). This leads to the increasing demand for online repositories of distributional word embeddings trained on different corpora in different languages and with different hyperparameters, like the NLPL Repository described by Fares et al. (2017).

The increasing popularity of machine-learned word embedding models caused the surge of research pushing the boundaries of existing methods and seeking to apply the same general idea to other types of input. Among others, one can mention Le and Mikolov (2014), who proposed Paragraph Vector, an algorithm to efficiently learn distributed representations not only for words but also for paragraphs or documents, and Bojanowski et al. (2017), who released fastText[11], a model able to learn embeddings using subword data (character n-grams), and thus partially solving the out-of-vocabulary (OOV) words problem.

[11] https://fasttext.cc/

More recently, in 2018, the field of distributional semantic modeling started paying attention to the so called 'contextualized word embeddings' which we already mentioned in the previous section. They provide different word representations in different contexts, unlike traditional 'static' embeddings where a single vector is attached to each single word. This comes at the cost of much higher computational requirements, but these requirements are not always perceived as problematic in the era of graphical processing units (GPUs) and Tensor Processing Units (TPUs) specifically designed for artificial neural network computations. Two widely acclaimed contextualized algorithms that advanced the state-of-the-art in many NLP tasks are:

- Embeddings from Language MOdels (ELMo), which use bidirectional Long Short-Term Memory (LSTMs) (Peters, Neumann, Iyyer, et al., 2018)
- Bidirectional Encoder Representations from Transformer (BERT), which use multi-layered transformers with attention (Devlin et al., 2019)

Models trained using such architectures can be used 'as is': contextualized representations are fed into the overarching system like the standard static embeddings, but this time the word vectors depend on the particular input context. Another mode of usage is that the whole model is fine-tuned on target task data (for example, sentiment analysis or natural language inference).

Interestingly, ELMo authors go further and claim that their architecture layers reflect language tiers, reflecting traditional structuring of language in linguistics (Peters, Neumann, Zettlemoyer, et al., 2018):

1. convolutional embedding layer reflects morphology;
2. the first LSTM layer reflects syntax;
3. the second LSTM layer reflects semantics (including word senses).

We describe contextualized embeddings in much more detail and apply ELMo and BERT to the task of diachronic semantic change modeling in Chapter 6 of the present thesis.

2.7 Other approaches to model meaning

The *data-driven* or distributional approach (including its word embedding variation) is significantly different from other methods to represent lexical meaning. It is impossible (and out of scope for this thesis) to cover all of them here, but we describe some examples below. In particular, computational semantics has long relied on *knowledge-driven* methods. They can be roughly categorized as follows:

- Dictionaries
- Ontologies

Dictionaries One natural (and arguably the oldest) way to represent word meaning is to simply use the lexicographical definitions from published dictionaries. The benefit here is that one can be sure of the representations quality: they are produced directly from human knowledge and are supposed to perfectly reflect the state of language semantics at a given moment. The obvious downside is that the textual definitions themselves are not machine-readable, and thus are not fit for large-scale NLP tasks (they have to be first converted to some numerical representations). Another problem is the inherent finite nature of any dictionary: it contains only as many words as the authors managed to process, and it is not easy to add new ones. This is especially important in the context of diachronic research. However, researchers still sometimes use this type of representations for semantic change detection (often together with other approaches); see, for example, R. Hu et al. (2019).

Ontologies Ontology-based approaches address some problems of dictionary-based ones, while still preserving the human-made quality. They consist of semantic networks (ontologies) or graphs relating human language words (or concepts) to each other. These networks as a rule are constructed manually by expert linguists. The most famous examples of such an ontology are, arguably, the WordNet project (Miller, 1995) and the BabelNet (Ehrmann et al., 2014). With typed relations between lexical entities presented in a machine -readable format, such resources are much easier to use in practical tasks. Unlike 'continuous' distributional representations, ontologies provide 'discrete' knowledge about word similarity or dissimilarity. They also naturally encode senses (by the possibility to explicitly link a word form to several different concepts or 'synsets'). However, building ontologies is still expensive and time-consuming, because of the manual work involved.

Dictionaries and ontologies represent knowledge-driven alternatives to word embeddings. But even within data-driven approaches there are models which differ significantly. For example, instead of aiming at finding representations for words, one can aim at representation for documents. We briefly mentioned topic modeling in Section 2.2; here we describe it as an alternative to techniques based on word embeddings.

Topic modeling This family of approaches is in fact strongly related to word embeddings, also being inherently distributional and vectorial. The Latent Semantic Analysis (Landauer and Dumais, 1997) described in Section 2.2 can be looked at as a topic modeling approach as well. However, the most popular algorithm in topic modeling is undoubtedly the Bayesian-based Latent Dirichlet allocation (LDA) (Blei, Ng, et al., 2003), with many different followers and variations. The common part here is that given a corpus of documents, these algorithms try to infer from the data a set of latent *topics*, of which each document is a mixture. The documents (essentially just word sequences) can be more or less similar to each other in terms of their topical structure: each document is represented with a vector of topic probabilities. Although these approaches are mostly focused on representing *documents*, vectorial representations of *single words* are also learned: these vectors reflect the probabilities of a particular word to occur in each of the inferred topics. Note that unlike dictionaries and ontologies (and like word embeddings), topic modeling is completely data-driven: given a corpus, one has to only specify the desired number of topics (algorithms like Hierarchical LDA allow one to avoid even this step and infer the number of topics automatically). From a high-level point of view, topic modeling and word embedding methods differ only in being focused on representing either documents or words correspondingly.

Note that distributional and knowledge-based approaches do not necessarily exclude each other. In fact, the latter can often help the former to overcome the lack of explicit linguistic 'competence' inherent to the data-driven paradigm. For example, one can use WordNet-like ontologies to improve data-driven word embeddings in the workflow known as 'retrofitting' (Faruqui, Dodge, et al., 2015). Dictionaries published at different time periods can be employed as the source of the ground truth for semantic change modeling systems which are themselves based on distributional vectors (Tsakalidis et al., 2019). In Chapter 6 of this thesis, the evaluation of embedding-based algorithms for semantic change detection is powered by the WordNet data.

As already mentioned in the Introduction, this thesis generally focuses on the methods which employ distributional word embeddings. One reason is that I am interested in this field, and it has been the primary focus of my research for several years. In addition, this choice is motivated by the success of word embedding-based approaches in many other semantic tasks and them being computationally efficient and easy to integrate into deep learning architectures.

2.8 Summary

To summarize, distributional word embeddings trained on large amounts of linguistic data efficiently capture many aspects of word meaning. As such, they are among the foundational bricks in the building of natural language processing systems able to at least partially 'understand' and generate human language. This is true, even when taking into account that the distributional hypothesis is more of an *assumption* and that word embeddings have a number of technical and conceptual shortcomings.

If word embeddings are able to infer word meaning at a given point in time, they provide a good starting point for research aimed at modeling semantic change automatically, in a 'data-driven' manner. Such representations form a strong empirical basis for linguistic hypotheses testing and may give answers to many questions regarding lexical semantic shifts.

The word embedding-based approaches are not the only existing methods for such modeling, and they were rarely used for it when the work on this thesis had started (back in 2015). However, since then, this family of approaches has definitely come to be the most popular in the field, as is clearly evidenced by the results of the first SemEval shared task in unsupervised lexical semantic change detection (Schlechtweg, McGillivray, et al., 2020). The overwhelming majority of the participants (including the best systems) used some variants of word embeddings. The quote from Tahmasebi, Borin, and Jatowt (2018) ‘The state of the art is represented by methods based on word embedding techniques’ seems to still hold in 2020.

In Chapter 3, we next survey the current state of using distributional representations for lexical semantic change modeling. Since distributional word embeddings are also the focus of the current thesis in general, it naturally pays more attention to the details of their usage. At the same time, other approaches to lexical semantic change modeling are also briefly described.

Chapter 3

Modeling diachronic semantic change: state of the field

In this chapter, we describe the current state of academic research relevant to our thesis. We begin by discussing the notion of 'semantic shift' itself, and then we continue with the history of attempts to model such shifts computationally.

As has already been mentioned before, this task is growing in popularity presently. There are dozens of papers on the topic, mostly published after 2011 (we survey some of them below). However, this emerging NLP field is still highly heterogeneous. There are at least three different research communities interested in it: natural language processing (and computational linguistics), information retrieval (and computer science in general), and social sciences.

These communities are not entirely isolated from each other, but are not strongly connected either. This is reflected in the terminology, which is far from being standardized. Most publications use distributional representations learned using neural networks in this or that form, but they can be referred to as 'temporal embeddings,' 'diachronic embeddings,' 'dynamic embeddings,' etc., depending on the background of a particular research group. Here is an example of how this can lead to misleading pointers. K et al. (2020) is a paper from the information retrieval community, and in it, the terms 'temporal embeddings' and 'dynamic embeddings' are used interchangeably, citing both Bamler and Mandt (2017) and Di Carlo et al. (2019). However, in fact, the joint training model from Bamler and Mandt (2017) and the 'temporal word embeddings with a compass' approach from Di Carlo et al. (2019) are entirely different, and K et al. (2020) use only the latter and this can confuse the reader. We note again that in this thesis, the term 'diachronic embeddings' is used in the sense of 'word embedding representations trained separately on time-specific corpora', while we reserve the term 'dynamic embeddings' for the models trained using the specific joint learning approach presented in Bamler and Mandt (2017) and Yao et al. (2018) (see subsection 3.2.5) or in Rudolph and Blei (2018). The 'temporal embeddings' expression does not seem to acquire wide recognition in the field, and we do not use it terminologically.

The year of 2018 saw several attempts to describe the diversity of approaches to semantic change detection, introduce some axes of comparison and outline main challenges which the practitioners face. Among such surveys, one can mention Kutuzov, Øvrelid, et al. (2018)[1], Tang (2018), Tahmasebi, Borin, and Jatowt (2018), and the Ph.D thesis by Dubossarsky (2018). In 2019, the first Workshop on Computational Approaches to Historical Language Change took place, collocated with the ACL conference (Tahmasebi, Borin, Jatowt, and Y.

[1] The current chapter is partially based on this paper.

Xu, 2019), and in 2020 the first SemEval shared task on unsupervised lexical semantic change detection was organized (Schlechtweg, McGillivray, et al., 2020). All these efforts are consolidating the field, which hopefully will help it to more strongly establish its presence in the wider NLP community.

In the following Section 3.1 we will discuss the notion of 'semantic shift' in linguistics.

3.1 Semantic shift as a linguistic concept

Human languages change over time, due to a variety of linguistic and non-linguistic factors and at all levels of linguistic analysis (Aitchison, 2001). In the field of theoretical diachronic linguistics, much attention has been devoted to expressing regularities of linguistic change. For instance, laws of phonological change have been formulated (e.g., Grimm's law or the great vowel shift) to account for changes in the linguistic sound system; in a similar vein, morphological and syntax changes have been analyzed (Hock and Joseph, 2019). When it comes to lexical semantics, linguists have long studied the evolution of word meaning over time, describing so-called lexical *semantic shifts* or *semantic change*. Bloomfield (1933) defines them as 'innovations which change the lexical meaning rather than the grammatical function of a form'. The general direction of a shift is $A \rightarrow B$, where A is the source meaning, and B is the target meaning (Zalizniak, 2018).

The central question here is of *semasiological* nature: what changes occurred to the meaning of a given lexeme, without considering changes to its form (Traugott, 1999). However, studies in diachronic *onomasiology* are also possible: a researcher is then interested in cases of lexical replacement, where one and the same meaning (concept) is expressed by different lexemes, as time passes by (Grzega and Schoener, 2007). Most of the current thesis is devoted to semasiological changes, except Section 5.3 in which we deal we the task of modeling diachronic changes in relations between words. This task can be looked at as onomasiological in some aspects.

Further we will review some of the main findings in the linguistic study of semantic shifts and relate these to methods currently employed in the field of Natural Language Processing (NLP).

Historically, much of the theoretical work on semantic shifts has been devoted to documenting and categorizing various types of shifts (Bréal, 1899; Stern, 1931; Bloomfield, 1933). The categorization found in Bloomfield (1933) is arguably the most used and has inspired a number of more recent studies (Blank, 1999; Geeraerts, 1997; Traugott and Dasher, 2001). Bloomfield (1933) originally proposed nine classes of semantic shifts, six of which are complimentary pairs along a dimension:

1. narrowing – broadening (widening);
2. hyperbole – meiosis;
3. elevation – degeneration;

4. metaphor;
5. metonymy;
6. synecdoche.

For instance, the pair 'narrowing – broadening' describes the observation that word meaning often changes to become either more specific or more general. In this way, Old English '*mete*' 'FOOD' becomes English '*meat*' 'EDIBLE FLESH', or the more general English word '*dog*' is derived from Middle English '*dogge*' which described a dog of a particular breed (Bloomfield, 1933). Bloomfield (1933) also describes change along the spectrum from positive to negative, describing the speaker's attitude as one of either degeneration or elevation, e.g. from Old English '*cniht*' 'BOY, SERVANT' to the more elevated '*knight*'.

The more current work of Geeraerts (1997) and Traugott and Dasher (2001) largely follows the categorization of Bloomfield, but focuses in particular on the processes of metaphorization and metonymization as driving forces in semantic shifts. Whereas metaphors are based on similarity, describing changes such as '*mouse*' meaning 'SMALL RODENT' being augmented with 'COMPUTER MANIPULATION DEVICE', metonymy is a usage such as '*drink a bottle*' where the container ('*bottle*') is being used to refer to its contents (Blank, 1999). Additionally, Geeraerts (1997) argued for the importance of encyclopedic information (or 'non-denotational meaning') to the study of semantic change; see the discussion of the distinction between different types of meaning in Chapter 2.

The driving forces of semantic change are varied, but include linguistic, psychological, social, cultural or encyclopedic causes (Blank and Koch, 1999; Grzega and Schoener, 2007). Linguistic processes that cause semantic change generally involve the interaction between words of the vocabulary and their meanings. This may be illustrated by the process of ellipsis, whereby the meaning of one word is transferred to a word with which it frequently co-occurs, or by the need for discrimination of synonyms caused by lexical borrowings from other languages. Semantic change (especially contextual variance) may also be caused by changes in the attitudes of speakers or in the general environment of the speakers.

Semantic shifts are naturally separated into two important classes: linguistic drift (slow and regular changes in core meaning of words, driven mostly by linguistic causes) and socio-cultural shifts (culturally determined changes in the people's associations of a given word). Socio-cultural semantic shifts are *changes in word meaning which are driven by non-linguistic exogenous social or cultural factors, for example, technological developments* (Hamilton, Leskovec, et al., 2016a). In the traditional classification by Stern (1931), cultural shifts correspond to the category of *substitution*. This may be exemplified by the word '*car*' which after the introduction of the automobile, changed its meaning from non-motorized vehicles to the new phenomenon. Changes in linguistic legislation (e.g. in the meanings of '*rape*' or '*harass*') is another example of an external, non-linguistic factor that influences lexical semantic shifts (Traugott, 2017). It should be noted that the boundary between linguistic and cultural shifts is not defined precisely, and many cases manifest features from both classes. However, the existence of this division have been shown empirically by Hamilton, Leskovec, et al. (2016a). To some extent, this division mirrors the general linguistic difference between functional and event-based triggers of language change (Bickel and Hickey, 2017).

Socio-cultural shifts can happen relatively quickly, unlike linguistically motivated ones, which typically may only be observed over decades or even centuries (Traugott and Dasher, 2001). Events considered to be important may immediately trigger substantial change in associations which comprise non-denotational meaning for a particular word. This is especially true for named entities which by their nature tend to associate with different concepts depending on what is happening around real-world phenomena which these entities denote. A good example is the word '*Kosovo*' acquiring a new aspect of meaning related to 'WAR' after the 1998-1999 military campaign. Note that this is not a 'semantic shift proper'. It is rather a change of usage, a typical attitude, or associations bound to this object in public opinion. Actually, this is one of reasons why sometimes researchers avoid the term 'meaning change' and refer to 'usage change' instead (Gonen et al., 2020).

However, within the current thesis we take the point of view that 'meaning' is actually determined by 'usage' (see Chapter 2 about the distributional hypothesis), and hence that their opposition is misleading. If 'senses' are clusters of word usages (Kilgarriff, 1997), then contextual variance (for example, '*Kosovo*' being used more in armed conflict contexts) can be thought of as change in the probability distribution of '*Kosovo*' senses. This is much related to the long-standing tradition in linguistics stating that meaning is a flux of relations in a situational context (Firth, 1935). Thus, we consider such cases to constitute semantic change as well.

The availability of large digital corpora have enabled the development of new methodologies for the study of semantic shifts within general linguistics (Traugott, 2017). A key assumption in much of this work is that changes in a word's collocational patterns reflect changes in word meaning (Hilpert, 2008), thus providing a usage-based account of semantics (Gries, 1999). For instance, Kerremans et al. (2010) studied the very recent neologism '*detweet*', showing the development of two separate usages/meanings for this word ('TO DELETE FROM TWITTER', vs 'TO AVOID TWEETING') based on large amounts of web-crawled data.

The usage-based view of lexical semantics is essentially the same as the assumptions underlying the distributional approach (see Chapter 2) often employed in NLP. In NLP research, the time spans studied are often considerably shorter (decades, rather than centuries) and we find that these distributional methods seem well suited for monitoring the gradual process of meaning change. Gulordava and Baroni (2011), for instance, showed that distributional models capture cultural shifts, like the word '*sleep*' acquiring more negative connotations related to sleep disorders, when comparing its 1960s contexts to its 1990s contexts. Here again, as mentioned previously, the core meaning of a word itself is not changed (there is no new sense of 'sleep'), but rather there are gradual changes in the contextual variance, signaling changes in speaker attitudes.

Hamilton, Leskovec, et al. (2016a) showed how it is possible to distinguish different driving forces in semantic change using different computational measures and emphasized the distinction between cultural shifts and linguistic drifts in natural language corpora (see more on this in Section 3.2.5). The current thesis often deals with cultural or associative shifts, represented by context variance.

To sum up, semantic change is often reflected in large corpora through fluctuations in the contexts of the word which is undergoing a shift, as measured by co-occurring words. Here, we are talking about semasiological shifts: people use other words more or less frequently together with the given word, because some aspects of the meaning of the given word have changed. In other cases, an onomasiological shift can happen, when another word X (or several new words) takes the position of a word Y which was used to denote some concept Z before. In fact, if X is not a neologism (and full neologisms are rare), then this phenomenon can be looked at both as onomasiological and as semasiological: on the one hand, the meaning Z stays the same, while its form changes from Y to Z, but on the other hand, in the course of this, the old meaning of X is obviously changing in some way. Thus, 'semasiological change' and 'onomasiological change' are rather two ways to look at the same process than two different types of processes.

Linguists today widely acknowledge that large-scale corpora can help understand language dynamics and change (Nölle et al., 2020). Thus, it is natural to try to detect semantic change automatically, in a data-driven way. In the following sections, we overview the methods currently used for unsupervised semantic change modeling and the recent academic research related to this problem.

3.2 Tracing semantic shifts distributionally

Conceptually, the task of discovery of diachronic semantic change from data can be formulated as follows. Given corpora $[C_1, C_2, ...C_n]$ containing texts created in time periods $[1, 2, ...n]$ correspondingly, the task is to find words with meaning changed between different time periods, or to rank words according to their level of meaning change. Other related tasks are possible: discovering general trends in semantic shifts (see Section 3.3) or tracing the dynamics of the relationships between words (see Section 3.4). In the next subsections, we address several axes along which one can categorize the research on detecting semantic shifts with distributional models.

3.2.1 Sources of diachronic data for training

When modeling semantic change in an unsupervised way, the types of generalizations we will be able to produce are much influenced by properties of the textual data being used, such as the sources and the temporal granularity of the corpora. In this subsection we discuss the data choices made by researchers.

The time unit (the granularity of the temporal dimension) can be chosen before slicing the text collection into sub-corpora. Earlier works dealt mainly with long-term semantic shifts (spanning decades or even centuries), since they are usually easier to trace. The early examples are Hilpert and Gries (2009) who studied frequency developments of words in the TIME corpus[2] and Sagi et al. (2009) who studied differences between Early Middle, Late Middle and Early Modern English, using the Helsinki Corpus (Rissanen et al., 1993).

A large role in further development of the field was played by the Google Books Ngrams corpus[3], which caused a surge of the new data-driven discipline of 'culturomics', studying human culture through digital media (Michel et al., 2011). Mihalcea and Nastase (2012) used this corpus to detect differences in word usage and meaning across 50-years time spans, while a bit earlier Gulordava and Baroni (2011) compared word meanings in the 1960s and in the 1990s, achieving good correlation with human judgments. Unfortunately, Google Ngrams is inherently limited in that it does not contain full texts (it is possible to download only 5-word fragments). However, for many cases, this corpus was enough and its usage as the source of diachronic data continued in Mitra et al. (2014), who detected word sense changes over decades.

In many of the following works, time spans decreased in size and became more granular. In general, corpora with smaller time spans are useful for analyzing socio-cultural semantic shifts, while corpora with longer spans are necessary for the study of linguistically motivated semantic shifts. As researchers are attempting to trace increasingly subtle cultural semantic shifts (often more relevant for practical tasks), the granularity of time spans is decreasing and the issue of *short-term semantic change* receives much attention. For example, Kim et al. (2014), Liao and Cheng (2016) and Del Tredici et al. (2019) analyzed yearly lexical changes.

In addition to the Google Ngrams corpus (with granularity of five years), Kulkarni et al. (2015) used Amazon Movie Reviews (with granularity of one year) and Twitter data (with granularity of one month). Their results indicated that computational methods for the detection of semantic shifts can be robustly applied to time spans less than a decade. Since then, Twitter data became a relatively popular choice for short-term semantic change modeling, with many datasets available, including the recently presented COVID-19 Twitter dataset containing about 14 billion word tokens (Banda et al., 2020). In a similar vein, Stewart et al. (2017) used the data from the Vkontakte social network to predict very short-term (up to several weeks) changes in semantic representations of words. Another popular and publicly available corpus for short-term diachronic studies is the Signal Media Dataset (Corney et al., 2016), which we employed in Kutuzov and Kuzmenko (2016).

Tahmasebi (2013) and Zhang et al. (2015) used the New-York Times Annotated Corpus (Sandhaus, 2008) with yearly sub-corpora, again managing to

[2] The TIME corpus contains about 275 000 articles from TIME magazine from 1923 to 2006, https://www.english-corpora.org/time/.

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

trace subtle semantic shifts. The same corpus was employed by Szymanski (2017), with 21 separate models, one for each year from 1987 to 2007, and to some extent by Yao et al. (2018), who crawled the New-York Times web site to get 27 yearly sub-corpora (from 1990 to 2016). Yao et al. (2018) captured semantic change with the granularity of years: for example, observing that the nearest neighbors for the proper noun '*Obama*' were moving from Barack Obama pre-presidential life in 1990-2006 ('*university*', '*professor*', '*civil*', etc) to political terms in 2008-2016 ('*president*', '*campaign*', '*government*', etc.), with the same trends observed for Donald Trump.

The inventory of diachronic corpora used in tracing semantic shifts was expanded by Jatowt and Duh (2014), who turned to the Corpus of Historical American (COHA)[4]. They used COHA as an additional source of data, with Google Ngrams being the main one. Hamilton, Leskovec, et al. (2016b) continued the usage of COHA along with the Google Ngrams corpus, and Eger and Mehler (2016) made the former their main data source (with the granularity of one decade). Cook, Lau, Rundell, et al. (2013) were the first to use two years of the English Gigaword news corpus (Parker et al., 2011), while in Kutuzov, Velldal, et al. (2017b), we employed all its yearly slices in the analysis of cultural semantic drift related to armed conflicts.

In Table 3.1 we list main English corpora which have been used for diachronic research with distributional approaches. The sizes of the corpora in word tokens are provided, but sheer size is not the only important property of a diachronic corpus. First of all, not all the corpora are publicly available: for example, New-York Times Annotated Corpus, COHA and Gigaword are available for a fee only, while Google Books Ngrams does not provide any clear ways to obtain the full corpus at all. Another aspect to consider is, of course, the time span covered by the corpus: the Helsinki Corpus might be small in comparison to Twitter or Gigaword, but if one is interested in Old English and Middle English, the latter corpora will not be of much help. Finally, the domain composition of the corpus can be of paramount importance: diachronic shifts occurring in movie reviews can be very different from those occurring in news pieces.

Note that Table 3.1 does not claim to be exhaustive. Nowadays, researchers start to use many other diachronic corpora in various languages besides English: the Deutsches Textarchiv, Berliner Zeitung and Neues Deutschland for German, the LatinISE for Latin, the Kubhist for Swedish, the Russian National Corpus and Lenta.ru dataset for Russian, the Corpus of Contemporary American English and Project Gutenberg for English, and many others, which would be impossible to list here. The CLARIN association maintains a list of historical corpora (mostly with long-term time spans) at https://www.clarin.eu/resource-families/historical-corpora.

[4] http://corpus.byu.edu/coha/

Corpus	**Size, words**	**Reference**
Helsinki Corpus	10^6	Rissanen et al. (1993)
New-York Times Annotated Corpus	$\approx 2 \times 10^9$	Sandhaus (2008)
Google Books Ngrams	$\approx 100 \times 10^9$	Michel et al. (2011)
English Gigaword	$\approx 4 \times 10^9$	Parker et al. (2011)
Corp. of Hist. Amer. Engl. (COHA)	400×10^6	Davies (2012)
Amazon Movie Reviews	$\approx 9 \times 10^8$	McAuley and Leskovec (2013)
Twitter (also in other languages)	$\approx 14 \times 10^9$	Banda et al. (2020)

Table 3.1: Popular English corpora for diachronic research

3.2.2 Evaluation of diachronic semantic change modeling

Diachronic datasets are needed not only as a source of *training* data for developing systems to trace semantic change, but also as a source of *test* sets to evaluate such systems. But in this case the situation is more complicated. Ideally, diachronic approaches should be evaluated on human-annotated lists of semantically changed words (preferably ranked by the degree of the shift). However, such gold standard data is difficult to obtain, even for English, let alone for other languages.

Works on language change from general linguistics like Traugott and Dasher (2001) or Daniel and Dobrushina (2016) and others as a rule contain only a small number of hand-picked examples, not enough to properly evaluate an automatic unsupervised system. The DatSemShift database (Zalizniak, 2018) features more than 4 000 semantic shifts across 800 languages. But it is focused on cognitive proximities between pairs of linguistic meanings (with a limited set of pre-defined senses): in this paradigm, a semantic shift is just a case of extended polysemy. The DatSemShift database is extremely useful for identifying recurring cross-linguistic semantic shifts, but it is yet to find out what is the best way to employ it for evaluation of unsupervised semantic change detection systems.

Thus, until recently, there were few standard test sets in the field, and the existing ones were of varying quality and availability. For example, Gulordava and Baroni (2011) manually annotated a dataset of English words by the degree of their semantic change from the 1960s to the 1990s (the GEMS dataset). Even though the inter-rater agreement was not high (see more on that in Chapter 6), this resource is still of enormous value for the field. However, the authors did not make the GEMS publicly available. Even eight years later, researchers have to contact the authors personally to get the dataset.

Fortunately, the situation starts to improve in the recent years. A prominent example is a package of test sets for English, German, Latin and Swedish provided in Schlechtweg, McGillivray, et al. (2020), accompanying the SemEval-

2020 shared task 1.[5] They are publicly available and manually annotated using a framework for the annotation of lexical semantic change called DURel or 'Diachronic usage relatedness' (Schlechtweg, Schulte im Walde, et al., 2018). The RuSemShift datasets for Russian by Rodina and Kutuzov (2020) follow the same approach, with certainly more to come in the nearest future. This standardization and unification of annotated test data is beneficial for the whole field.

Typically, such test sets are simply lists of words where each word is accompanied either by a binary class label (where '1' means 'semantic shift' and '0' means 'no shift') or by a continuous value representing the degree of semantic change. The list is associated with two different time spans (for example, the 19th century and the 20th century) and the corresponding corpora of texts produced within these time spans. An automatic system is supposed to predict the class label or the change score. The first case corresponds to the task of binary semantic change classification and the second case corresponds to the task of estimating and quantifying the degree of semantic change (we mentioned these two main aspects of diachronic semantic change modeling in the Introduction). As a rule, the classification predictions are evaluated with accuracy or F-1 score, while the change scores are evaluated with the Spearman rank correlation between the predictions of the system and human annotations.

Unfortunately, manually annotated semantic change datasets are still unavailable for the majority of world languages, and those that are available are rather small. Doubts are expressed, for example, about whether one can trust Spearman rank correlations calculated on sets of 30 or 40 elements (Gonen et al., 2020). Thus, the problem of evaluating approaches to semantic change modeling is far from being solved, and practitioners often rely on self-created test sets, or even on simple eyeballing of the results.

Various ways of overcoming this problem without extensive manual annotation have been proposed. For example, Mihalcea and Nastase (2012) evaluated the ability of a system to detect the time span that specific contexts of a word undergoing a shift belong to ('word epoch disambiguation'). A similar problem was offered as SemEval-2015 Task 7: 'Diachronic Text Evaluation', where the participants were challenged to automatically determine the period when a text was written (Popescu and Strapparava, 2015). Another possible evaluation method is the so-called 'cross-time alignment', where a system has to find equivalents for certain words in different time periods (for example, '*Obama*' in 2015 corresponds to '*Trump*' in 2017). Arguably, manual annotation for such datasets is easier to obtain than for full-fledged semantic change datasets. There exist several test sets containing such temporal equivalents at least for English (Yao et al., 2018).

Another interesting direction is the usage of the existing dictionaries or thesauri which contain the year when a particular word sense had been introduced. This approach is taken in the dataset presented in Cook, Lau, McCarthy, et al. (2014) based on Macmillan English Dictionary for Advanced Learners

[5] We work with these test sets extensively in Chapter 6.

(MEDAL), and in the datasets introduced by Tahmasebi and Risse (2017a) and Tsakalidis et al. (2019) and based mostly on the Oxford English Dictionary. In the same vein, Aggelen et al. (2019) presented the large HiT dataset based on the Historical Thesaurus of English[6]. Notably, they also released several prior work datasets in a unified format, which is going to be very helpful in further evaluation and comparison efforts. However, high quality dictionaries (especially ones which contain diachronic sense information) are still a scarce resource for the majority of world languages.

Yet another evaluation strategy is to use the computed diachronic semantic change to trace or predict real-world events like armed conflicts, which took place in the corresponding time spans. Thus, event datasets (created and annotated by researchers in other fields of science: history, political studies, social studies, etc.) can serve as proxies to language change. A somewhat similar idea was employed in Wijaya and Yeniterzi (2011), who checked that the periods of the detected semantic shifts coincide with political events in these time spans. However, they did not develop it into a full-fledged evaluation framework. We employed this approach in Kutuzov, Velldal, et al. (2017a), Kutuzov, Velldal, et al. (2017b), Kutuzov, Velldal, et al. (2019), and in this thesis in Chapter 5.

Finally, when lacking manually annotated datasets of semantic shifts, one can turn to so called ‘synthetic evaluation’. It is rooted in the field of word sense disambiguation (WSD), where artificially created ‘ambiguous’ pseudo-words have long been used to evaluate supervised algorithms (Schütze, 1998). In WSD, pseudo-words are injected in real corpora to imitate synchronic lexical polysemy. In semantic change modeling, such pseudo-words are injected to imitate polysemy changing diachronically (for example, a word gradually acquiring or losing a sense over time). Since these words are injected by a researcher and known by definition, the gold standard data emerges naturally. Synthetic evaluation was applied to semantic shift detection by Dubossarsky, Hengchen, et al. (2019) and Shoemark et al. (2019), among others. However, it should always be kept in mind that synthetic data follows a researcher’s assumptions about how real semantic shifts should behave. It is never the same as real annotated data, and thus the conclusions drawn from synthetic evaluation should be taken with a grain of salt.

3.2.3 Pre-embedding approaches to semantic change modeling

After settling on a diachronic data set to be used in the system (both for training and for testing), one has to choose which data-driven methods to employ. Since our task belongs within the field of semantics, this implies the choice of a particular type of meaning representation. As already discussed in Chapter 2, the spectrum of existing data-driven (‘distributional’) representations of meaning is very rich. They can be document-centric or word-centric. If the meaning is expressed by vectors, these vectors can be sparse and explicit (with interpretable components) or dense and distributed (with non-interpretable components); also,

[6] https://ht.ac.uk/

these vectors can be produced directly from a co-occurrence matrix ('count-based') or trained via optimizing language modeling loss ('prediction-based'). Practically any combination of these representation types can be employed for semantic change modeling. This thesis is focused on a particular combination: trained dense word vectors, also known as 'word embeddings' (Baroni et al., 2014). But first, in this subsection, we will outline prior work which uses other corpus-based approaches

A word can be represented with its corpus frequency only: in fact, at some time point it was quite common to use change in raw word frequencies in order to trace semantic shifts or other kinds of linguistic change. For examples of such work see, among others, Juola (2003); Hilpert and Gries (2009); Michel et al. (2011); Lijffijt et al. (2012); Bochkarev et al. (2014), or Choi and Varian (2012) for frequency analysis of words in web search queries. Naturally, frequency-based methods can be useful in detecting the emergence of neologisms (Ryskina et al., 2020)[7] or the disappearance of existing words, which is arguably more important for onomasiological research (Tjong Kim Sang, 2016), while vector representations are useless if there is no co-occurrence data to infer them from. In fact, a large part of 'culturomics' (Michel et al., 2011) revolves around using frequencies to trace the introduction of new entities into common usage by language speakers (or how some entities are going into oblivion). The algorithm here can be as simple as calculating the absolute or normalized difference between target word frequencies in two time-specific corpora.

However, if one wants to trace semasiological changes to existing word form, then using raw frequency differences obviously has its limitations. Semantic shifts are not always accompanied with strong changes in word frequency (or this connection may be very subtle and non-direct). Since words belong to different frequency tiers, and absolute frequency values are not distributed across the vocabulary uniformly, it is difficult to find a robust method to calculate frequency differences between diachronic corpora. Nowadays, raw frequency is as a rule used only as the simplest possible baseline for semantic change detection systems (Schlechtweg, McGillivray, et al., 2020).

As a sort of transfer towards full-scale distributional representations, researchers also studied the increase or decrease in the frequency of a word A collocating with another word B over time, and made conclusions about changes in the meaning of A (Heyer, Holz, et al., 2009). Although these collocates had to be manually defined prior to any experimentation, this allowed the researchers to capture phenomena like the English '*web*' collocating mostly with '*spider*' in 1994, but mostly with '*designer*' in 2014, corresponding to the emergence of the new 'INTERNET' sense (McEnery et al., 2019). However, naively representing lexical semantics through specific word collocates suffers from the lack of generalization power. At the same time, extending the list of collocates to the entire vocabulary (as in Berberich et al. (2009) start suffering from the curse of dimensionality (see Chapter 2 for the discussion of this issue in the context of explicit distributional

[7] In addition to frequencies, this work also uses semantic sparsity information inferred from word embeddings.

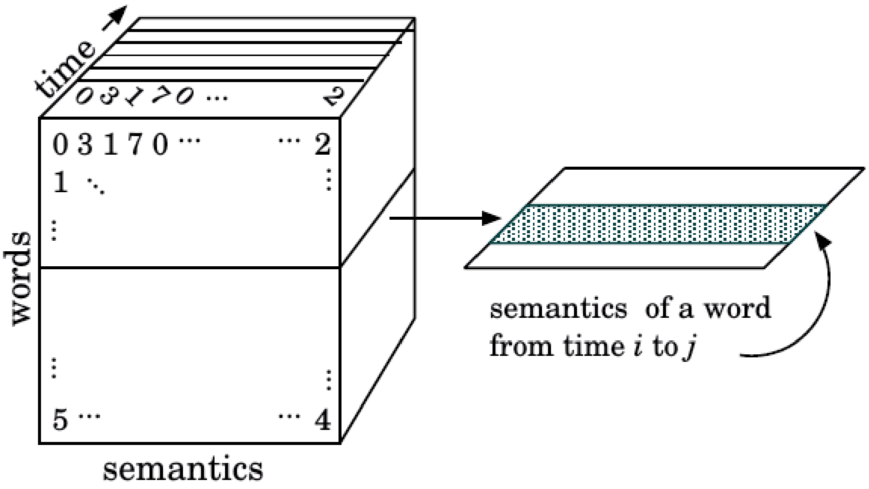


Figure 3.1: Tensor representation of a semantic space; image from (Jurgens and Stevens, 2009).

models). There were also attempts to trace diachronic semantic change through the shifts in grammatical relations of target words (Gerow and Ahmad, 2012), but they didn't lead to large-scale success either.

Around 2009, it was proposed that one can use vector-based distributional methods (similar to modern ones) to reliably detect semantic shifts which are not manifested through frequency change or simple collocates change. The pioneering work by Jurgens and Stevens (2009) described an insightful conceptualization of a sequence of distributional representations changing through time: it is effectively a $Word \times SemanticVector \times Time$ tensor, in the sense that each word possesses a set of semantic vectors for each time span we are interested in. The more different are the time-specific vectors, the higher is the degree of semantic change between the corresponding time bins.

This concept is graphically represented in Figure 3.1. It paved the way for quantitatively comparing not only words with regard to their synchronic meaning, but also different stages in the development of word meaning over time. This conceptualization still remains the foundation of the whole field of using distributional representations to diachronic semantic change modeling.

Jurgens and Stevens (2009) employed the Random Indexing (RI) algorithm (Kanerva et al., 2000) to create word vectors from a training corpus, while Sagi et al. (2009) turned to Latent Semantic Analysis (Deerwester et al., 1990). Both these methods already worked with dense vectors: technically, the only difference between these representations and modern word embeddings was that they were not trained via language modeling. However, the 'word2vec revolution' was still several years ahead, and two years later Gulordava and Baroni (2011) still used explicit representations consisting of sparse word co-occurrence matrices weighted by Local Mutual Information, with a similar approach taken by Tahmasebi, Gossen, et al. (2012), who traced named entity evolution.

Basile, Caputo, et al. (2014) proposed an extension to Random Indexing, dubbed *Temporal Random Indexing.* No quantitative evaluation of this approach was offered (only a few hand-picked examples based on the Italian texts from the Gutenberg Project), and thus it is unclear whether Temporal Random Indexing is any better than other distributional models for the task of semantic shift detection. A newer study by Basile and McGillivray (2018) does evaluate Temporal Random Indexing but lacks comparison to modern word embedding algorithms.

Further on, the diversity of the employed methods increased, with graph approaches gaining popularity. For example, Mitra et al. (2014) analyzed clusters of the word similarity graph in the sub-corpora corresponding to different time periods. Their distributional model consisted of lexical nodes in the graphs connected with weighted edges. The weights corresponded to the number of shared most salient syntactic dependency contexts,where saliency was determined by co-occurrence counts scaled by Mutual Information (MI). Importantly, they were able to detect not only the mere fact of a semantic shift, but also its type: the birth of a new sense, splitting of an old sense into several new ones, or merging of several senses into one. Thus, this work goes into a much less represented class of 'fine-grained' approaches to semantic shift detection. Other examples of graph-based approaches are Tahmasebi (2013) and Tahmasebi and Risse (2017a) who tracked individual sense changes (word sense evolution) on the basis of the curvature clustering algorithm. In these works, the concept of word sense differentiation is of great importance. Our analysis of word senses in the context of diachronic semantic change modeling with word embeddings can be found in Chapter 2 and Chapter 6.

Another vein of research focused on sense changes employed topic modeling approaches (where topics are interpreted as senses). Prominent example is Lau et al. (2012) who applied LDA in conjunction with non-parametric Hierarchical Dirichlet Process. Senses were naturally mapped to automatically inferred corpus topics, so that the distribution of word senses corresponds to its topic probabilities. The paper is largely devoted to the task of word sense induction (WSI), but then the same technique is used to find words which acquired a novel sense over time. Interestingly, their approach is token-based (each word occurrence receives its own sense distribution, handling polysemy and contextual variance naturally) which makes it somewhat similar to contextualized token embeddings we employ in Chapter 6. With this, Lau et al. (2012) managed to distinguish lemmas with a novel sense from semantically stable 'distractor lemmas' better than the frequency baseline. However, due to the lack of proper semantic change datasets, they had to rely on a very small self-created test set containing only five shifted and five stable English lemmas. Thus, this work was rather exploratory.

Cook, Lau, McCarthy, et al. (2014) improved the evaluation of novel sense detection task by presenting the manually annotated SiBol/Port English test set[8] containing 13 lemmas gaining a new sense between 1993 and 2010. They also extended the method from Lau et al. (2012) by taking into account the relevance of each induced sense, which is calculated based on keywords of the current corpus relative to the previous time-specific corpus. The intuition here is that it can be helpful to know for what topics we expect to see novel senses (for example, 'computing' is a relevant topic when comparing 2010 to 1993). However, even with 13 lemmas, the proposed dataset was rather small and the results on it were not much better than the frequency baseline.

Accordingly, we believe that non-parametric topic modeling approaches have great potential for semantic change detection. The vein of research based on *dynamic topic modeling* (Blei and Lafferty, 2006; Wang and McCallum, 2006), which learns the evolution of topics over time, is rather strong. In Wijaya and Yeniterzi (2011), it helped solve a typical digital humanities task of finding traces of real-world events in the texts. Heyer, Kantner, et al. (2016) employed topic analysis to trace the so-called 'context volatility' of words. Frermann and Lapata (2016) drew on these ideas to trace diachronic word senses development (we compare our embeddings-based approach to theirs in Chapter 6). In the political science, topic models are also sometimes used as proxies to social trends developing over time: for example, Mueller and Rauh (2017) employed Latent Dirichlet Allocation (LDA) to predict timing of civil wars and armed conflicts.

But most scholars nowadays seem to prefer parametric distributional models, particularly prediction-based embedding algorithms like SGNS, CBOW or GloVe. We outline a surge of word embedding-based research in the next subsection.

3.2.4 Embedding approaches to semantic change modeling

Word embeddings are dense and continuous vector representations of lexical semantics trained in an iterative unsupervised fashion with the target to improve loss on the language modeling task. Typical examples are architectures like word2vec (Mikolov, Sutskever, et al., 2013) and fastText (Bojanowski et al., 2017). Following their widespread adoption in NLP in general, they have become the dominant representations for the analysis of semantic change as well.

We emphasize again that the word embedding-based approaches are not the only existing methods for semantic change modeling. However, this family of methods is now the most widely used in the field, as is clearly evidenced by the results of the first SemEval shared task in unsupervised lexical semantic change detection (Schlechtweg, McGillivray, et al., 2020): 18 of 21 participants (including all the winners) used either static or contextualized word embeddings. We believe this provides additional justification to our focus on this type of semantic representations, apart from us simply being interested in what diachronic information they can capture.

As discussed in Chapter 2, dense distributional representations provide an efficient way to tackle synchronous semantic tasks. They represent lexical meaning with dense vectors (embeddings), produced from word co-occurrence

[8] Unfortunately, the URLs from the paper for this test set are not valid any more.

counts. Although conceptually the source of the data for these representations is still word and collocate frequencies, unlike count-based methods, they 'compress' this information into continuous vector representations which are both efficient and convenient to work with (Baroni et al., 2014).

The work of Kim et al. (2014) was seminal in the sense that it is arguably the first to employ prediction-based word embedding models to trace diachronic semantic shifts. Particularly, they used Continuous Skipgram with negative sampling (SGNS) (Mikolov, K. Chen, et al., 2013).[9] Along with that, they introduced the incremental or chronological training approach (see subsection 3.2.5 below), leveraging new properties of prediction-based embeddings. Kim et al. (2014) successfully identified semantic shifts in widely used examples like the English word '*cell*' (the beginning of the 21 century). Already back then, they understood the limitation of such method in that it cannot determine the nature of the shift (narrowing or widening, amelioration or pejoration, etc). This concern is still valid as of now.

Kulkarni et al. (2015) empirically demonstrated that distributional word embeddings outperform the frequency-based methods in modeling diachronic semantic shifts. They managed to trace semantic change more precisely and with greater explanatory power. One of the well-known examples from their work is the semantic evolution of the word '*gay*': through time, its nearest semantic neighbors were changing, manifesting the gradual move away from the sense of 'CHEERFUL' to the sense of 'HOMOSEXUAL' (see Figure 3.2).

Hamilton, Leskovec, et al. (2016b) showed the superiority of SGNS over explicit PPMI-based distributional models in semantic change modeling, although they noted that low-rank approximations of explicit models with singular value decomposition (SVD) (Bullinaria and J. P. Levy, 2007) can perform on par with SGNS, especially on smaller datasets. Since then, the majority of publications in the field started using dense word representations: either in the form of SVD-factorized PPMI matrices, or in the form of prediction-based shallow neural models like SGNS[10].

Embedding models provide dense vector representations which are both efficient, scalable and very convenient to integrate into NLP pipelines, including those for semantic change detection. However, apart from general shortcomings of word embeddings, described earlier in Chapter 2, there are some issues specifically related to their usage in semantic change modeling. Arguably, the most important is the problem of making the embedding models comparable (sometimes called 'the problem of alignment'), which we discuss in the next subsection 3.2.5.

3.2.5 Comparing embeddings across time

As already mentioned in Chapter 2, it is not straightforward to compare vector representations across different separately trained embedding models, even if

[9] Continuous Bag-of-Words (CBOW) from the same paper is another popular choice.

[10] We remind that O. Levy and Goldberg (2014) showed these two approaches to be equivalent from the mathematical point of view.

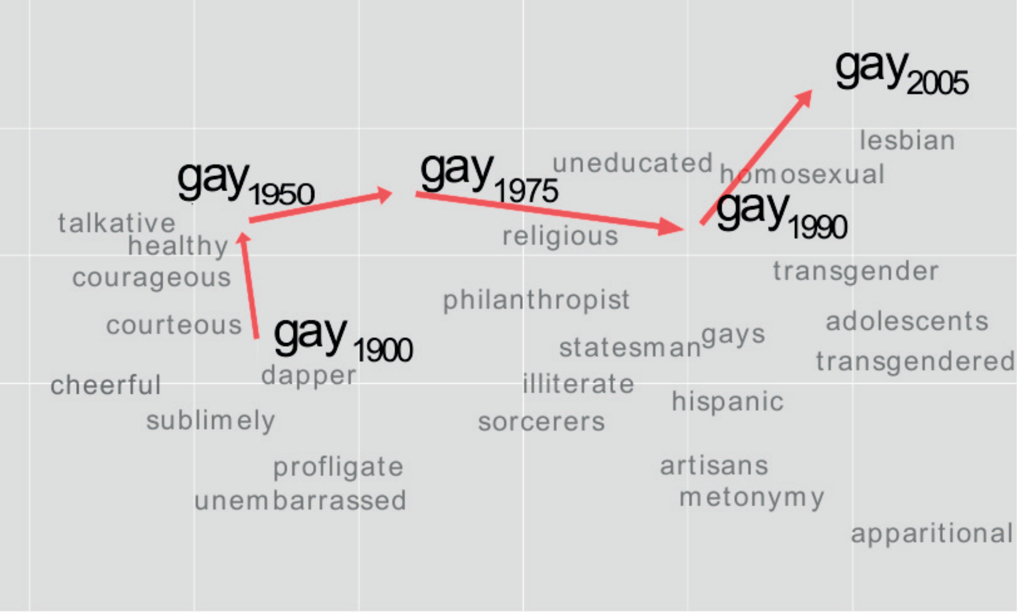


Figure 3.2: Semantic trajectory of the English word '*gay*' in the space of its context words (Kulkarni et al., 2015).

their vocabulary is essentially the same (diachronic word embedding). The reason is the non-deterministic and stochastic nature of the prediction-based embeddings. The most popular remedy here is to *align* different vector spaces by somehow making the vectors comparable: but this is not the only one. In this subsection, we describe approaches to overcome this problem.

First of all, it is entirely possible to discard the global states of vector spaces under comparison and look only at the *local* lists of k nearest neighbors produced by different diachronic word embeddings for one and the same target word (where k is much smaller than the size of full vocabulary). One can then estimate similarity of these lists, for example, using Jaccard similarity coefficient (Jaccard, 1901) or Kendall's τ (Kendall, 1948). The lower the similarity, the stronger is the semantic shift (if the lists are entirely different, the word meaning is entirely changed). In doing this, we are moving from the vector space to the word space, and naturally, the nearest words to a particular word will be more or less the same in all runs of the same training algorithm on the same corpus (provided they are trained for long enough): they do not depend on random initialization like the values of particular vector components. Thus, we can expect that semantically stable words will have similar nearest neighbors in embedding models trained on diachronic corpora. Another benefit of this method is its immediate interpretability: it is much easier for a human to understand the nature of a semantic shift by looking at two lists of, say, 10 nearest words than by looking at a single cosine similarity score. This approach was used to trace semantic shifts between different domains or corpora in Kutuzov and

Kuzmenko (2015) and Gonen et al. (2020), among others. However, it has significant downsides: if considering only a few nearest neighbors (in the order of dozens), the method becomes too *local*: it might miss the cases when a word is moving in the semantic space *together with its neighbors*. On the other hand, increasing the number of considered neighbors quickly becomes computationally expensive and sensitive to random fluctuations. Also the nearest neighbor-based approaches do not take into account the relations between the neighbors themselves. Overall, directly comparing local 'semantic neighborhood' remains an unpopular technique: the participants of the recent SemEval shared task on unsupervised lexical semantic change detection did not use it at all (Schlechtweg, McGillivray, et al., 2020). Still, we demonstrate an example of employing this approach below in Chapter 4.

If one wants to consider the entire vector spaces when modeling semantic change, one has to first make the embeddings trained on different corpora comparable (that is, vectors for semantically similar stable words trained on C_1 should yield high cosine similarity to those trained on C_2). In one of the early publications, Kulkarni et al. (2015) suggested *aligning* the models to fit them in one vector space, using linear transformations preserving general vector space structure. The idea is as follows. If we are given two independently trained embedding matrices A and B with a significant shared vocabulary (for example, trained on two diachronic corpora), we can find an orthogonal linear transformation T such that it projects A to B while minimizing the squared loss. This problem, shown in in Equation 3.1, is solved using the Orthogonal Procrustes method (Gower, Dijksterhuis, et al., 2004), which is also popular in the field of cross-lingual word embeddings (Artetxe et al., 2020).

$$T = \underset{T}{\text{argmin}} ||T \cdot A - B||^2 \tag{3.1}$$

After A is projected to the B vector space, cosine similarities between their vectors become meaningful and can be used as indicators of semantic change. Kulkarni et al. (2015) also proposed constructing the time series of a word embedding over time, which allows for the detection of 'bursts' in its meaning with the *Mean Shift* model (Taylor, 2000). Notably, almost simultaneously the idea of aligning diachronic word embedding models using a distance-preserving projection technique was proposed by Zhang et al. (2015). They were dealing with the temporal correspondence problem in which, given a query term and the source time period, the task is to find the counterpart of the query that existed in the target time period (see Section 3.4). Zhang et al. (2016) expanded on this by adding the so called 'local anchors': that is, they used both linear projections for the whole models and small sets of nearest neighbors for mapping the target words to their correct temporal counterparts. Tsakalidis et al. (2019) showed that it can also be beneficial to base the Procrustes transformation on a limited set of diachronically stable 'anchor words'. These approaches operate both in the space of vector representations and in the space of particular words.

Direct alignment with orthogonal projections is easy and straightforward to use. This approach is sometimes criticized for its self-contradicting objective (it attempts to project each word to itself, even in the presence of a shift) and for instability with respect to different embedding spaces (Gonen et al., 2020). However, it is often very efficient in semantic change detection, as shown by Shoemark et al. (2019) and by us in Chapter 6 of the present thesis. And in case its performance is not satisfactory, there is a number of alternatives, described below

Instead of aligning their diachronic embedding using linear transformations, Eger and Mehler (2016) compared word meaning using so-called 'second-order embeddings': that is, the vectors of words' cosine similarities to *all* other words in the shared vocabulary of all models. This approach does not require any alignment at all: basically, one simply analyzes the word's position compared to other words. The absolute values of cosine similarities in two models under analysis will almost never be the same (because of different vector space density), but this is not important for this method. What is important is the relative ranking of the words from the shared vocabulary by the similarity to the target word. If the ranking is more or less the same, the similarities' vectors from both models will be very similar themselves (by dot product or cosine similarity), and the conclusion would be that the word semantics has not changed. If, vice versa, the rankings are substantially different, the similarities' vectors will yield low cosine similarity between themselves, leading to the conclusion that a semantic shift occurs.

A very similar algorithm was described by Yin et al. (2018) under the name of 'Global Anchors' (meaning that all the words from the vocabulary are used as anchors). Hamilton, Leskovec, et al. (2016b) and Hamilton, Leskovec, et al. (2016a) showed that these two approaches can be used simultaneously: they employed both 'second order embeddings' and linear transformations to align diachronic models. Interestingly, J. Xu et al. (2019) used these methods of vector space alignment not only for language-related tasks, but also for exploring temporal patterns in dynamic graphs of any nature (they call it 'diachronic node embeddings'). Note that although the 'second-order embeddings' or Global Anchors technically operate in the word space, they are very different from the nearest neighbor comparison approach: they take into account the full global structure of the vector space, and their results are not directly interpretable (instead of two short word lists, one has two high-dimensional second-order similarity vectors). Also, one can argue that computing the intersection of two models' vocabularies already constitutes a sort of 'alignment'.

Further on, it was shown in Bamler and Mandt (2017) (dynamic skip-gram model) and Yao et al. (2018) (dynamic Word2Vec model, DW2V) that it is possible to learn word embeddings across several time periods jointly, enforcing comparability across all of them simultaneously, and positioning all the representations in the same vector space in one step. This eliminates the need to first learn separate embeddings for each time period, and then align each subsequent model pair. The method in Bamler and Mandt (2017) is conceptually similar to dynamic topic models (Blei and Lafferty, 2006; Rudolph and Blei, 2018) and combines a Bayesian version of the word2vec Skipgram architecture with a latent time series. They additionally describe two variations of their approach, for the cases when data slices: a) arrive sequentially, as in streaming applications, and one can not use future observations; b) are available all at once, allowing for training on the whole sequence from the very beginning. Note, however, that the dynamic skip-gram architecture assumes some pre-existing division of the training corpus into specific documents (this is the consequence of being based on the topic modeling idea), which is not the case for classic word embeddings.

'Dynamic word2vec' from Yao et al. (2018) avoids this assumption and learns time-aware embeddings by jointly optimizing embeddings for all time periods in question. In this case, a form of 'alignment' is enforced through regularization which smooths changes across time. The main advantage of this method is that it can be easily used on any number of time bins, while being robust against cases when a particular time bin has less data and thus yields embeddings of lower quality. Note that despite the name of the method mentioning 'word2vec', in fact it employs sparse co-occurrence count matrices weighted by positive pointwise mutual information (PPMI) and factorized to reduce their dimensionality. This does not change the results: as we already mentioned in Chapter 2, O. Levy and Goldberg (2014) showed that the objective of word2vec is equivalent to low-rank factorization of a PPMI matrix. However, it does influence the computational performance of the method: materializing all these matrices can require a prohibitive amount of RAM. The authors propose to solve this issue with scalable block coordinate descent, but unfortunately, do not provide any reference code to re-implement their approach. Further on, an interesting extension was presented by Rosenfeld and Erk (2018) who train a deep language modeling network with word and time representations. Word vectors in this setup are learned linear transformations applied to a continuous time variable, and thus producing an embedding of word w at time t.

Dubossarsky, Hengchen, et al. (2019) proposed Temporal Referencing, which also alleviates the need to explicitly align the vector spaces, while still making it possible to calculate cosine similarity between word embeddings from different time periods. One embedding model is trained on all n time bins combined, with n time-specific vectors learned for each target word. This is done by replacing each time-agnostic target word token w with a time-specific token w_n at train time: for example, in the corpus corresponding to the 1970s, '*computer*' becomes '*computer_1970*' when it is a target word, but stays as it is when it is a context word. As a result, time-specific target word representations are naturally located in a shared vector space. Dubossarsky, Hengchen, et al. (2019) report increased performance of SGNS embeddings with Temporal Referencing in comparison to SGNS with alignment: mainly because their method is better in understanding that a word is semantically stable (less susceptible to random noise). Dubossarsky, Hengchen, et al. (2019) empirically evaluated Temporal Referencing only on the Word Sense Change test set by Tahmasebi and Risse (2017b), which contains 13 changed and 19 stable English words, and the authors themselves acknowledge that 'the results are indicative rather than conclusive'.[11]

[11] Dubossarsky, Hengchen, et al. (2019) also conducted evaluation using a synthetic change dataset, which supported the superiority of Temporal Referencing. However, as we already

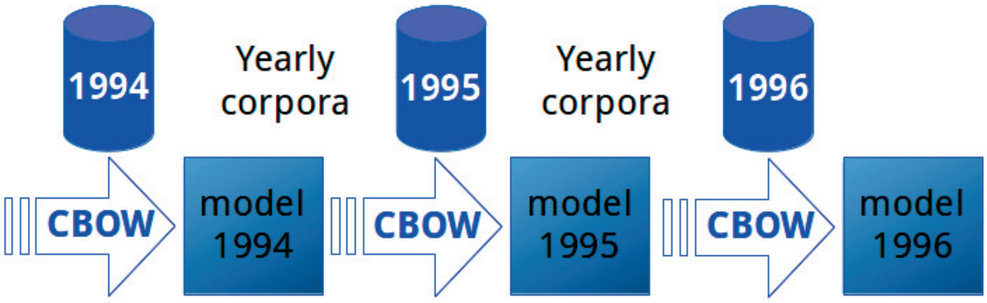


Figure 3.3: Incremental training of word embeddings (in this example, CBOW models are trained on yearly diachronic corpora).

In fact, Schlechtweg, Hätty, et al. (2019) had the opposite results (they use the term 'Word Injection' instead of 'Temporal Referencing'). They evaluated on the German DURel semantic shift dataset which was annotated following the de-facto standard semantic change annotation framework (Schlechtweg, Schulte im Walde, et al., 2018), and found that Orthogonal Procrustes alignment consistently outperforms Temporal Referencing. The size of DURel is comparable to the Word Sense Change test set. Contradicting evaluation results can be explained by different tasks which were being solved: Dubossarsky, Hengchen, et al. (2019) dealt with the classification task trying to distinguish stable words from those that were changed, while Schlechtweg, Hätty, et al. (2019) dealt with the ranking task trying to order words by the degree of their semantic change. It is possible that Procrustes alignment and Temporal Referencing can be complimentary to each other, depending on the type of the task. Also note that Temporal Referencing requires knowing what target words are going to be analyzed *before actually training the embedding model*, which may not always be possible in practice. In theory, it is possible to replace *all* words with their time-specific tokens; but this will mean the n-times explosion of the vocabulary size.

Yet another way to make diachronic embeddings comparable is made possible by the fact that prediction-based word embedding approaches (as well as Random Indexing) allow one to update the trained models with new data. This is not the case for the explicit count-based algorithms, which usually require a computationally expensive dimensionality reduction step. Kim et al. (2014) proposed the idea of *incrementally updated* diachronic embeddings: that is, they train a model on the year y_i, and then the model for the year y_{i+1} is initialized with the word vectors from y_i. See Figure 3.3 for a schema of this workflow. It is also known as 'vector initialization' or VI (Schlechtweg, Hätty, et al., 2019).

mentioned, any evaluation of semantic change modeling on synthetic data should be taken with a grain of salt.

This is also an alternative to post-hoc alignment: instead of aligning models trained from scratch on different time periods, one starts with training a model on the diachronically first period, and then updates this same model with the data from the successive time periods, saving its state each time. Thus, all the models are inherently related to each other. This, again, makes it possible to directly calculate cosine similarities between the same word in different time-specific embeddings, or at least makes the models much more comparable. The method is extremely straightforward and easy to implement using off-the-shelf libraries. Dubossarsky, Weinshall, et al. (2016) used such incrementally updated embeddings to compare the speed of semantic change for different parts of speech.

Several works aim to address the technical issues accompanying this approach of incremental updating. Among others, Peng et al. (2017) described a novel method of incrementally learning the hierarchical softmax function for the Continuous Bag of Words and Continuous Skipgram algorithms. In this way, one can update word embedding models with new data and new vocabulary much more efficiently, achieving faster training than when doing it from scratch, while at the same time preserving comparable performance. Continuing this line of research, Kaji and Kobayashi (2017) proposed a conceptually similar incremental extension for negative sampling, which is a method of training examples selection, widely used with prediction-based models as a faster replacement for hierarchical softmax.

Unfortunately, as far as we know, these techniques were not continued in further works. Partially this may be due to the fact that incrementally trained diachronic embeddings are often sub-optimal in comparison to aligned ones (Shoemark et al., 2019). One of the non-obvious issues here is the one of vocabulary extension: as the model is trained with additional data, there should be some procedure to add new lexical entries to its vocabulary, based on their frequency. It is extremely difficult to do it right and to avoid either catastrophic forgetting of the past or insensitivity to the present. The same is true for choosing the vector dimensionalities and the optimal number of updating epochs: it highly depends on word frequencies (Schlechtweg, Hätty, et al., 2019; Kaiser et al., 2020). Thus, incremental training should be used with caution; in this thesis we demonstrate both the cases when it is beneficial (Chapter 5) and cases when it loses to linear projection based alignment (Chapter 6).

Finally, the relatively novel 'contextualized embedding' architectures like ELMo (Peters, Neumann, Iyyer, et al., 2018) or BERT (Devlin et al., 2019) offer an entirely new approach to the problem of making diachronic embeddings comparable. Unlike the 'static' embedding models like word2vec, fastText, GloVe, etc, that produce *type embeddings*, they work with *token embeddings*: that is, context-dependent representations of words. It means that a contextualized model can be pre-trained on all the time bins of the diachronic corpus concatenated (or even on an entirely different large corpus in the same language). After that, this model can be used to infer token embeddings for each occurrence of a target word in time-specific corpora; recall that similar conceptualizations were used by Lau et al. (2012) with a topic modeling approach to semantic change detection. These token embeddings will be similar for tokens used in similar contexts and different for tokens used in different contexts. This makes it possible to formulate various ways of estimating the difference of token embeddings between two or more time-specific corpora, without the need to explicitly align anything: the embeddings are produced by one and the same language model and are comparable by design. This approach to semantic change modeling is relatively new, but was already successfully employed in several works (R. Hu et al., 2019; Giulianelli et al., 2020; Martinc, Montariol, et al., 2020; Martinc, Kralj Novak, et al., 2020). We discuss the advantages and downsides of using contextualized embeddings for diachronic studies in Chapter 6.

It was already mentioned that different methods of comparing vector representations are more or less 'global' or 'local'. The distinction between *global* and *local* embedding comparison methods was first introduced by Hamilton, Leskovec, et al. (2016b) and Hamilton, Leskovec, et al. (2016a), who made an important observation that it is correlated with the distinction between linguistic and cultural semantic shifts. The global methods take into account the whole model (for example, simple cosine similarity between two aligned vectors or 'second-order embeddings', when we compare the word's similarities to all other words in the lexicon), while the local methods focus on the word's immediate neighborhood (for example, when comparing the lists of k nearest neighbors). They concluded that global measures are more sensitive to regular processes of linguistic shifts, while local measures are better suited to detect chaotic cultural shifts in word meaning and usage. Thus, the choice of particular embedding comparison approach should depend on what type of semantic change one seeks to detect.

3.3 Laws of semantic change

The use of diachronic word embeddings for studying the dynamics of word meaning has resulted in several hypothesized 'laws' of semantic change. We review some of these law-like generalizations below, before finally describing a study that questions their validity.

Dubossarsky, Tsvetkov, et al. (2015) experimented with K-means clustering applied to SGNS embeddings trained for evenly sized yearly samples for the period 1850–2009. They found that the degree of semantic change for a given word (quantified as the change in self-similarity over time) negatively correlates with its distance to the centroid of its cluster. This distance to the centroid is also known as 'prototypicality': the degree to which a word is representative of the category of which it is a member of. For example, if considering the category of pets, '*cat*' is arguably a more prototypical pet than '*axolotl*' (although pet axolotls do exist). In a good word embedding model, the distance of the '*cat*' vector to the centroid of the '*pet*' cluster will be much lower than the same distance for the '*axolotl*' vector. It was proposed that the likelihood of semantic change correlates with the degree of prototypicality (the 'law of prototypicality' in Dubossarsky, Weinshall, et al. (2017)).

Another relevant study is reported by Eger and Mehler (2016), based on two different graph models; one being a time-series architecture relating embeddings across time periods to model semantic shifts and the other modeling the self-similarity of words across time. Experiments were performed with time-indexed historical corpora of English, German and Latin, using time-periods corresponding to decades, years and centuries, respectively. To enable comparison of embeddings across time, second-order embeddings encoding similarities to other words were used, as described in 3.2.5, limited to the core vocabulary (words occurring at least 100 times in all time periods). Based on linear relationships observed in the graphs, Eger and Mehler (2016) postulate two laws of semantic change:

1. a word embedding can be expressed as a linear combination of its neighbors in previous time periods;
2. the meanings of words tend to decay linearly in time, as modeled in terms of the similarity of a word to itself; this is in line with the 'law of differentiation' proposed by Y. Xu and Kemp (2015).

In another study, Hamilton, Leskovec, et al. (2016b) considered historical corpora for English, German, French and Chinese, spanning 200 years and using time spans of decades. The goal was to investigate the role of frequency and polysemy with respect to semantic shifts. As in Eger and Mehler (2016), the rate of semantic change was quantified by self-similarity across time-points (with words represented by Procrustes-aligned count-based SVD embeddings). Through a regression analysis, Hamilton, Leskovec, et al. (2016b) investigated how the change rates correlate with frequency and polysemy, and proposed another two laws:

1. frequent words change more slowly ('the law of conformity');
2. polysemous words (controlled for frequency) change more quickly ('the law of innovation').

Azarbonyad et al. (2017) postulated that these laws (at least the law of conformity) hold not only for time-specific corpora, but also for other 'viewpoints'. For example, semantic shifts can be observed across embeddings trained on texts produced by different political actors or written in different genres (Kutuzov, Kuzmenko, and Marakasova, 2016).

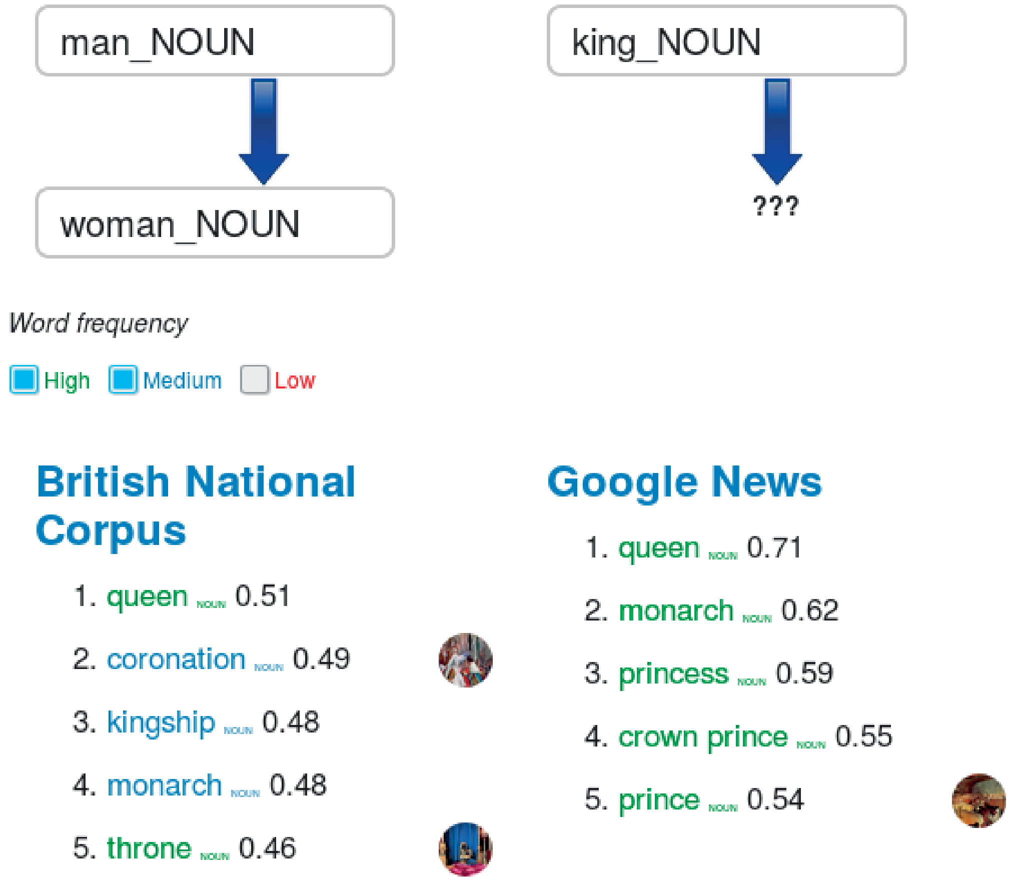
In principle, this leads to significant expansion of the 'semantic change' notion itself. Conceptually, the same techniques that are used for diachronic semantic change detection can be used for analyzing cross-domain semantic differences. In our previous work, we used Russian word embedding models to trace the difference of lexical meanings between a general-purpose corpus and a web-corpus (Kutuzov and Kuzmenko, 2015). Later, Schlechtweg, Hätty, et al. (2019) evaluated some of the aforementioned techniques on German synchronic semantic shift test set *SUReL*, described in Hätty et al. (2019). Gonen et al. (2020) re-invented the nearest neighbor comparison technique employed in Kutuzov and Kuzmenko (2015) to find the words with shifted usage in texts produced by authors of differing ages, genders, occupations or even published in different days of week. However, in this thesis we stick to the 'temporal' aspect of semantic change. Note that this allows for a view of the corpora under analysis as an ordered sequence of more than two elements, which is impossible in the cross-domain setup. It also increases the difficulty of the task.

Interestingly, Dubossarsky, Weinshall, et al. (2017) questioned the validity of some of these proposed laws of semantic change. In a series of replication and control experiments, they demonstrated that some of the regularities observed in previous studies are largely artifacts of the models used and frequency effects. Dubossarsky, Weinshall, et al. (2017) considered 10-year bins comprising equally sized yearly samples from Google Books 5-grams of English fiction for the period 1990–1999. For control experiments, they constructed two additional data sets; one with chronologically shuffled data where each bin contains data from all decades evenly distributed, and one synchronous variant containing repeated random samples from the year 1999 alone. Any measured semantic shifts within these two alternative data sets would have to be due to random sampling noise.

Then, Dubossarsky, Weinshall, et al. (2017) performed experiments using raw co-occurrence counts, PPMI weighted counts, and SVD transformations (Procrustes aligned), and conclude that the 'laws' proposed in previous studies – that semantic change is correlated with frequency, polysemy (Hamilton, Leskovec, et al., 2016b) and prototypicality (Dubossarsky, Tsvetkov, et al., 2015) – are not entirely valid as they are also observed in the control conditions. They suggested that these spurious effects are instead due to the type of word representation used (count vectors) and that semantic shifts must be explained by a more diverse set of factors than distributional ones alone. Dubossarsky, Weinshall, et al. (2017) did not use trained prediction-based embeddings, but in the following paper by Dubossarsky, Hengchen, et al. (2019), SGNS embeddings were shown to contain noise as well. In particular, SGNS-based methods can falsely detect 'semantic change' in stable control words which increased their frequency. Note, however, that Dubossarsky, Hengchen, et al. (2019) did not specifically focus on proving or disproving any laws. Thus, the discussion on the existence of the 'laws of semantic change' manifested by distributional trends is still open.

3.4 Diachronic semantic relations

Distributional methods can be used not only to trace meaning drift for particular single words. Word embeddings are known to successfully capture complex *relationships* between concepts, as manifested in the well-known word analogies task (Mikolov, K. Chen, et al., 2013). An example of this is presented in Figure 3.4, where the distributional model captures the fact that the relation between '*man*' and '*woman*' is the same as between '*king*' and '*queen*'. If one adds the '*king*' vector to the '*woman*' vector and subtracts the '*man*' vector, the resulting



man_NOUN

woman_NOUN

king_NOUN

???

Word frequency

■ High ■ Medium □ Low

British National Corpus

1. queen NOUN 0.51
2. coronation NOUN 0.49
3. kingship NOUN 0.48
4. monarch NOUN 0.48
5. throne NOUN 0.46

Google News

1. queen NOUN 0.71
2. monarch NOUN 0.62
3. princess NOUN 0.59
4. crown prince NOUN 0.55
5. prince NOUN 0.54

Figure 3.4: Analogy task in a vector space, with the results from two distributional models (WebVectors service (Fares et al., 2017)).

vector will have ‘*queen*’ as the closest word in the models’ vocabulary. Thus, it is a natural development to investigate whether changes in semantic relationships (links between the meanings of words) across time can also be traced by looking at diachronic embeddings.

The task of finding ‘temporal co-references’ (Tahmasebi, Gossen, et al., 2012) is to identify the word in a target time period which corresponds to a query term in the source time period (for example, given the query term ‘*iPod*’, the counterpart term in the 1980s time period is ‘*Walkman*’). Such identification is supposed to improve the results of information retrieval from document collections with significant time depth (Berberich et al., 2009). Tahmasebi, Gossen, et al. (2012) addressed this problem using plain co-occurrence graphs, while Zhang et al. (2015) called the same phenomenon ‘temporal correspondences’ and employed prediction-based Skipgram word embeddings. Note that it is natural to think about temporal correspondences in terms of *onomasiological* change (over time, another word form emerges to express the same concept), unlike the examples in the previous sections, which mostly focused on *semasiological* processes (over time, another concept emerges to be expressed by the same word form).

Szymanski (2017) framed this as the *temporal word analogy* problem, extending the word analogies concept into the temporal dimension. They showed that diachronic word embeddings can successfully model relations of the type 'word w_1 at time period t_α is like word w_2 at time period t_β'. They aligned embeddings trained on different time periods using linear transformations (see Section 3.2.5). Then, the temporal analogies were solved by simply finding out which word vector in the time period t_β is the closest to the vector of w_1 in the time period t_α. Later, Yao et al. (2018) solved the same problem using dynamic embeddings without alignment. A variation of this task was studied in Rosin et al. (2017), where the authors learn the relatedness of words over time, for the practical purpose of helping information retrieval, answering queries like 'in which time period were the words '*Obama*' and '*president*' maximally related'. This technique can be used for a more efficient user query expansion in general-purpose search engines.

In our work, we model a different semantic relation: 'words w_1 and w_2 at time period t_α are in the same semantic relation as words w_3 and w_4 at time period t_β' (Kutuzov, Velldal, et al., 2017a) . To trace the temporal dynamics of these relations, we re-apply linear projections learned on sets of w_1 and w_2 pairs from the model for the period t to the model trained on the subsequent time period $t + 1$. This is used to address the task of detecting lasting or emerging armed conflicts and the armed groups involved in these conflicts (we talk more about this in Section 5.3). Orlikowski et al. (2018) employed a similar framework to analyze diachronic evolution of concepts on a corpus of Dutch newspapers from the 1950s and the 1980s. This last paper is again a perfect example of onomasiological research: the studied semantic shifts are cases of diachronic lexical replacement (Tahmasebi, Borin, and Jatowt, 2018), where different words are coming into use to express one and the same concept.

Additionally, there exists a whole bulk of studies devoted to the issues of extracting *synchronic* semantic relations from word embeddings models. D. Chen et al. (2017) pointed out that some relations are predicted much better that others. They argue that the reason is that vector space approaches cannot model relations that violate symmetry or triangle inequality. For example, humans judge '*North Korea*' to be more similar to '*China*' than the other way around; this is a violation of symmetry. In another study, Gábor et al. (2017) also describe problems with predicting semantic relations, rooted in their symmetry and selectional restrictions on the answers. They propose taking into account the second order similarity in order to alleviate these problems. In their experiments, this considerably improved the task of unsupervised relation classification. Such problems are directly linked to the task of temporal semantic relations modeling, and it is crucial to closely study the nature of the semantic relations that one is trying to trace diachronically. At the same time, the amount of training data (the size of corpora) is as a rule much larger for the synchronic semantic relation tasks, so they are somewhat easier than the diachronic ones.

3.5 Applications

Practical applications of diachronic word embedding-based algorithms can generally be grouped into two broad categories:

1. *linguistic studies* which investigate the how and why of semantic shifts,
2. *event detection* setups which mine text data for actionable purposes.

The first category generally involves corpora with longer time depth, since linguistic changes happen at a relatively slow pace. Some examples falling into this category include tracking semantic drift of particular words (Kulkarni et al., 2015), identifying the breakpoints between epochs (Sagi et al., 2011; Mihalcea and Nastase, 2012), studying the laws of semantic change at scale (Hamilton, Leskovec, et al., 2016a) and finding different words with similar meanings at different points in time (Szymanski, 2017). Chapters 4 and 6 of the present thesis also fit into this category.

Diachronic studies have been held up as good use case of deep learning for research in computational linguistics (Manning, 2015). There are opportunities for future work applying distributional diachronic representations not only in the field of historical linguistics, but also in related areas like socio-linguistics and digital humanities. A good example here is the JeSeMe web service[12] described in Hellrich et al. (2018), which allows one to analyze temporal dynamics of emotions related to a particular word. The emotions are described via three scales:

1. valence (positive – negative),
2. arousal (calm – excited),
3. dominance (controlled – in control).

Figure 3.5 taken from the said web service shows how the English word '*climate*' starts evoking more negative emotions in the 1990s, while at the same time the level of excitement around this topic rises: this is obviously related to the public discussion about global climate change.

Word Evolution[13] (Jatowt, Campos, et al., 2018) is another web service based on Google Ngrams and the Corpus of Historical American English. It allows a user to trace changes in semantic associates of the query words across time. We created a conceptually similar ShiftRy web service[14] for analyzing short-term semantic change in Russian news texts, with extended visualization capabilities (Kutuzov, Fomin, et al., 2020). It fulfills the need for more user-friendly 'story-telling' systems for diachronic semantic change analysis.

ShiftRy makes a smooth transition to the second category of applications involving mining texts for cultural semantic shifts (usually on shorter time spans)

[12] http://jeseme.org/

[13] https://www.okayama.silk.jp/WordEvolution/

[14] https://shiftry.rusvectores.org/en/

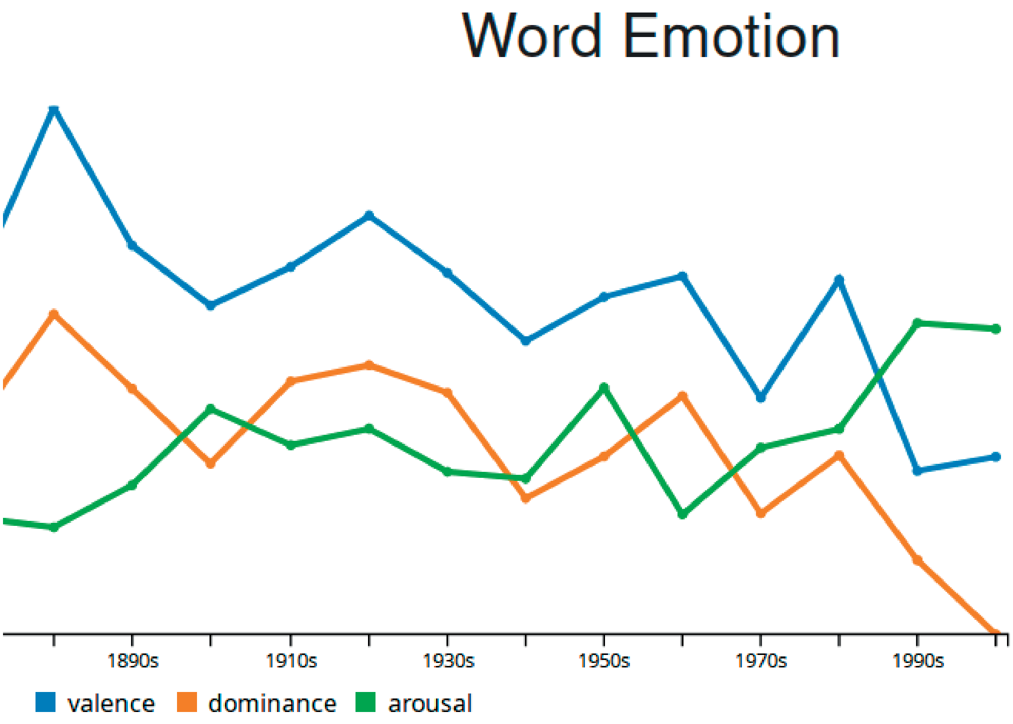


Figure 3.5: Diachronic changes of emotions associated with the English word '*climate*' as visualized by JeSeMe (Hellrich et al., 2018). The vertical axis (labels omitted for readability) reflects the values of emotion scales, with 0 in the middle.

indicating real-world events. Examples of this category are predicting civil turmoils like in Chapter 5 of this thesis or in Kutuzov, Velldal, et al. (2017b) and Mueller and Rauh (2017), temporal information retrieval in Rosin et al. (2017), or tracing the popularity of entities using norms of word vectors in Yao et al. (2018). Such systems can be employed to improve user experience in search engines or for policy-making in governmental structures.

We believe that the near future will see a more diverse landscape of applications for unsupervised semantic change modeling, especially related to the real-time analysis of large-scale news streams. 'Between the lines', these data sources contain a tremendous amount of information about processes in our world, manifested in semantic shifts of various sorts. The task of researchers is to reveal this information and make it reliable and practically useful.

3.6 Summary

In this chapter, we have presented an outline of the current research related to computational modeling of semantic change. We covered the linguistic nature of semantic shifts (both semasiological and onomasiological), the typical sources

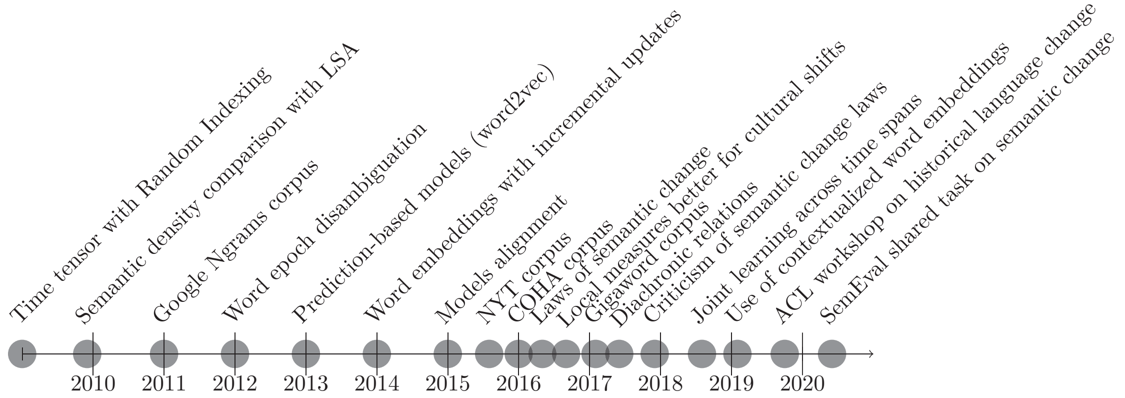

Time tensor with Random Indexing
Semantic density comparison with LSA
Google Ngrams corpus
Word epoch disambiguation
Prediction-based models (word2vec)
Word embeddings with incremental updates
Models alignment
NYT corpus
COHA corpus
Laws of semantic change
Local measures better for cultural shifts
Gigaword corpus
Diachronic relations
Criticism of semantic change laws
Joint learning across time spans
Use of contextualized word embeddings
ACL workshop on historical language change
SemEval shared task on semantic change
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020

Figure 3.6: Modeling diachronic semantic change distributionally: research timeline

of diachronic data for training and testing, and the distributional approaches used to model it: from frequency-based methods to static and contextualized word embeddings. This emerging field is still relatively new, and although recent years has seen a string of significant discoveries and academic interchange, much of the research still appears slightly fragmented. This chapter is partly aimed at addressing this issue and presenting computational detection of diachronic semantic shifts with word embeddings as a coherent story.

Figure 3.6 shows the timeline of the development of research in the area of unsupervised semantic change modeling: introducing concepts, usage of corpora, important findings and community events.

The next chapters of the present thesis employ several of the embedding-based approaches described in this chapter. In particular, in Chapter 4, we apply both global (including Global Anchors) and local (including the nearest neighbors comparison) methods for semantic change speed estimation. Chapter 5 employs a local method to detect diachronic changes in connotational meaning associated with armed conflicts (Section 5.2), while global methods are employed to trace drift of semantic relations (Section 5.3). Chapter 6 deals with semantic shifts proper and uses only global methods. To evaluate the abilities of diachronic embeddings to capture information about semantic change, we use real-world event datasets (Chapter 5) and manually annotated temporal semantic shift test sets (Chapter 6).

As for the approaches to make embeddings comparable, in Chapter 4 we either do not do this at all (since it is not needed for Global Anchors and nearest neighbor comparison) or use Orthogonal Procrustes alignment. For the less mainstream armed conflict related tasks presented in Chapter 5, incremental training is shown to outperform Orthogonal Procrustes. However, for the standard semantic change detection workflow described in Chapter 6, the Procrustes baseline is much stronger than the incremental training baseline. Still, they both are outperformed by techniques based on contextualized embeddings, which can be seen as another approach to make diachronic representations comparable. Contextualized architectures also provide rich possibilities for analysis, while at the same time having some intrinsic issues to be taken into account (see Chapter 6 for details). But in the next two chapters 4 and 5 we limit ourselves to standard 'static' embeddings.

Chapter 4

Measuring diachronic evolution of evaluative adjectives

In Chapter 3, we suggested that quantitative information about diachronic semantic change extracted from word embeddings can be used in linguistic case studies. In this chapter, we conduct research in exactly this vein, measuring the intensity of diachronic semantic shifts in evaluative adjectives in English, Norwegian and Russian across five decades.[1]

'Evaluative' adjectives are defined as those which describe object qualities from the subjective point of view of the speakers, expressing their opinions about the object being described. Typical English examples are '*good*', '*bad*' or '*brilliant*'. We test a particular linguistic intuition: that evaluative adjectives are more prone to diachronic semantic change than other types of adjectives (that shifts in their meaning are more probable and occur faster).

4.1 Motivation

Although we are not aware of any publication which explicitly claims that evaluative adjectives change faster, mentions of evaluative words (including adjectives) being prone to semantic change abound in scientific works.

We already discussed the categorization of semantic shifts. Borkowska and Kleparski (2007) studied semantic shift categories naturally related to evaluative words: namely, amelioration (acquiring more positive sentiment) and pejoration (acquiring more negative sentiment). They found these types to be extremely strong and wide-spread. In Hamilton, Clark, et al. (2016), the authors induced historical sentiment lexicons from English corpora (using word embeddings, among other methods). They showed that amelioration and pejoration do occur on a massive scale: many evaluative adjectives in English have completely switched their sentiment during the last 150 years (probably due to their emotional load). Multiple examples of evaluative words changing their polarity can be found in Traugott and Dasher (2001).

Thus, we know that there exist many cases of evaluative adjectives shifting their sentiment. It does not in itself mean that evaluative adjectives are specifically important from the viewpoint of modeling diachronic semantic change. In principle, any lexical cluster can be studied using the same methods: see, for example, Dubossarsky, Weinshall, et al. (2016) comparing the change degrees of verbs, nouns and adjectives in general. Evaluative adjectives present a group of words which is well-defined linguistically, and for which it is comparatively easy to obtain word lists. Also, adjectives do not receive as much attention

[1] Parts of this chapter were previously published as Rodina, Bakshandaeva, et al. (2019).

as nouns in the semantic change detection field. This was our motivation to choose this particular lexical cluster. We argue that this is exactly the case where quantitative techniques do provide evidence to general linguistics (not exhaustive evidence, of course).

Let us consider the English words '*incredible*' and '*terrific*' which underwent amelioration and started to denote positive instead of negative qualities. The online version of the Merriam-Webster dictionary[2] states that the word '*incredible*' was first used in the negative sense of 'TOO EXTRAORDINARY AND IMPROBABLE TO BE BELIEVED', but nowadays it also has a positive sense of 'AMAZING, EXTRAORDINARY'. Analogously, the original sense of '*terrific*' was 'VERY BAD; FRIGHTFUL'; today the first sense is 'UNUSUALLY FINE; MAGNIFICENT'. Note that both words still retain their original negative senses, but these senses seem to become less central in their use. An example of pejoration is the word '*pathetic*': it moved on from the sense of 'PASSIONATE' to the much more negative sense of 'PITIFULLY INFERIOR OR INADEQUATE'.

On the other hand, regular adjectives can also *become evaluative* in the course of semantic shifts happening across time: consider the history of the English word '*monumental*' from the 1960s to the 2000s (Figure 4.1 shows a t-SNE projection of its nearest distributional neighbors changing over time) or how the word '*sick*' slowly acquires a colloquial evaluative meaning ('*That's sick, dude!*'), as described in Mitra et al. (2014).

So, the examples of evaluative adjectives changing their sentiment (and thus, their semantics) are multiple. But do these examples stem from sheer hand-picking, or is there a general trend in human languages which makes evaluative adjectives more prone to some types of semantic shift over time? To answer this question, six different methods of quantifying semantic change are applied in this chapter. We extend prior work by studying not only sentiment changes, but semantic shift in evaluative adjectives in general. Additionally, we analyze data from three languages (English, Norwegian and Russian), and focus on a more narrow time span. Our time period is limited to only the decades from 1960s to 2000s. The reason for this is the availability of substantial amounts of reliable textual data for all three languages. As a sanity check, we also conduct additional experiments on data that spans 10 decades for English.

Frequency-controlled experimental results show that, depending on the particular method, evaluative adjectives either do not differ from other types of adjectives in terms of semantic change or appear to actually be less prone to it. Thus, in spite of many well-known examples of semantically changing evaluative adjectives, it seems that these processes are not particularly characteristic of this specific type of words: at least with relatively short-term time spans (on the order of several decades). Our experiments also show some limitations of word embedding-based methods (see Section 4.5).

[2] https://www.merriam-webster.com/dictionary/

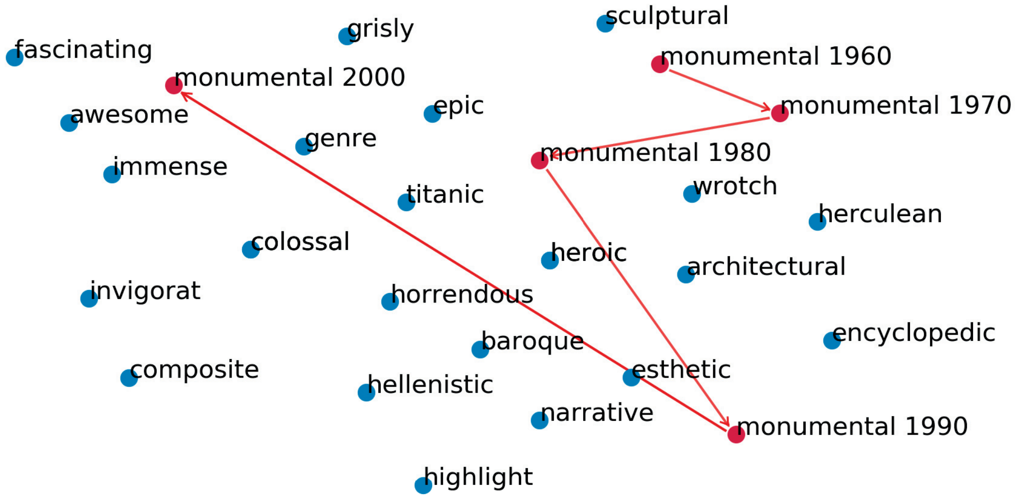


Figure 4.1: Alterations in the nearest distributional neighbors of the English adjective '*monumental*': from '*sculpture*' in the 1960s to '*awesome*' in the 2000s. t-SNE projection of CBOW vectors trained on the COHA corpus.

4.2 Training corpora and evaluative lexicons

In the following section, we present the diachronic corpora used in our experiments with evaluative adjectives, as well as the word embedding models trained on these corpora. Additionally, the process of creating evaluative adjective lexicons for three languages is described. Note that here we are interested in semantic change as a continuous multiple-point process, not as the difference between two time bins. For this reason, we needed diachronic corpora spanning across several consecutive time periods.

4.2.1 Corpora

For the purposes of this experiment, we employed corpora in three languages, selecting texts which were created during the five decades from the 1960s to the 2000s:

- For the English data, we used The Corpus of Historical American English (COHA).[3] We remind the reader that this is a corpus of English texts annotated with creation dates and balanced by genres. It is composed of fiction, magazine and newspaper articles, as well as non-fiction texts.

[3] https://www.english-corpora.org/coha/

	English	**Norwegian**	**Russian**
1960s	12.0	6.0	10.0
1970s	12.0	21.0	10.0
1980s	13.0	25.5	9.0
1990s	14.5	40.5	20.0
2000s	15.0	21.0	39.5

Table 4.1: Corpora sizes (in millions of words).

- For Norwegian data, we used the NBdigital corpus.[4] It contains texts in Norwegian Bokmål from the National Library of Norway's collection of texts in public domain. They are mainly documents produced by various public institutions. The texts have been digitized and converted to a machine readable form (OCR-recognized) automatically; for each text, the average OCR confidence is preserved. We kept only the texts with the OCR confidence higher than 0.9, to exclude poorly recognized cases.
- For Russian data, we used the Russian National Corpus (RNC).[5] It includes a wide variety of genres of written and spoken language, such as non-translated works of fiction, memoirs, essays, journalistic works, scientific and popular scientific literature, public speeches, letters, diaries, documents, etc. It is important that the RNC is also rigorously balanced across genres and types of texts.

The NBdigital corpus was provided to us already lemmatized and POS-tagged with the Oslo-Bergen tagger (Johannessen et al., 2012), along with syntactic disambiguation. The English and Russian corpora were lemmatized and POS-tagged by ourselves, using the corresponding UDPipe 2.3 models (Straka and Straková, 2017). Lemmatization was especially important for Russian with its rich morphology (at least 18 syntactic forms for each adjective).

Table 4.1 lists the corpora sizes for each decade and language under consideration. We also conducted additional experiments with the English texts covering a longer COHA time span: all the decades from the 1910s to the 2000s. The size of each of the pre-1960s sub-corpora is similar to the post-1960s ones: about 12 or 13 million word tokens. Unfortunately, we were not able to collect comparable (in size and reliability) diachronic corpora for Norwegian and Russian. Thus, the additional experiments with semantic shifts on a longer period of time are limited to English only. This dataset is referred to below as 'English10', since it covers 10 decades.

[4] https://www.nb.no/sprakbanken/ressurskatalog/oai-nb-no-sbr-43/

[5] http://ruscorpora.ru/en/

4.2.2 Word embeddings

Continuous bag-of-words embeddings (Mikolov, Sutskever, et al., 2013) were trained from scratch on each decade's sub-corpus for each of the three languages. We did not need to train the models incrementally, since all the methods we use in this chapter (see section 4.3) either employ Orthogonal Procrustes (which aligns the models itself) or work on the level of the nearest words (which avoids the need to make the embeddings comparable). Our prior work with lexical semantic change detection for Russian (Fomin et al., 2019) (not included in this thesis) showed that for semantic shifts proper, alignment based methods outperform incremental training based ones, which is in line with the results for other languages (Shoemark et al., 2019; Schlechtweg, Hätty, et al., 2019).

All the models share the same set of hyperparameters: vector size 300, symmetric context window size 3, and 10 training iterations (epochs) over the corpus. We discarded all the words which occurred less than five times in the training corpus, and additionally limited the maximum vocabulary size to be 100 000, so the less frequent words ranked below 100 000 were discarded as well (this way, we ensured comparability of the model vocabularies' sizes). The pre-trained embedding models are publicly available via the NLPL word vector repository[6] (Fares et al., 2017).

4.2.3 Evaluative adjective lexicons

In order to find out whether evaluative adjectives are more prone to diachronic semantic change, we need an authoritative source providing us with a list of such adjectives, preferably with a large number of them. Unfortunately, even for English such a list is hard to find in the published works, and the same is true for Norwegian and Russian. For this reason, we turned to sentiment lexicons: lists of positive and negative words widely used in natural language processing for the purposes of automatic sentiment analysis. The reason behind this choice was that such words are almost always evaluative by definition. Below we describe these lexicons for each of the three languages under analysis.

The lists for English and Norwegian come from the same source. The English list is a general sentiment lexicon composed of a positive and a negative part. These were created by assigning the positive and negative labels using a WordNet-based bootstrapping approach (M. Hu and B. Liu, 2004).[7] We thereafter automatically translated (from English to Norwegian) these positive and negative sentiment lexicons. The translations were manually checked, and corrected when necessary. Furthermore, if an English word had several senses that could be translated into different Norwegian words, these were added to the translations. We have omitted all multi-word expressions, and only kept single word translations. This resulted in a collection of 3 961 negative and 1 646 positive Norwegian words. The original English lexicons contained 4 783 negative and 2 006 positive words. We did not investigate rigorously to what

[6] http://vectors.nlpl.eu/repository/

[7] Available at https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Language	**Source**	**Total entries**
English	Customer Review Dataset (M. Hu and B. Liu, 2004)	2 250
Norwegian	same as English (translated)	1 939
Russian	RuSentiLex (Loukachevitch and Levchik, 2016)	2 435

Table 4.2: Evaluative adjective lexicons.

extent the translated lexicon is representative of the Norwegian language, but we believe that it is representative enough, since it is a general lexicon equivalent to its original English counterpart, and because the Norwegian list was checked manually to filter out non-evaluative adjectives.

The Norwegian lexical resource SCARRIE[8], a full-form lexicon, was used to identify which of the Norwegian translations were adjectives. Once these Norwegian adjectives were identified, we selected only the English words that had a Norwegian adjective as translation. Subsequently, we used the WordNet (Miller, 1995) to identify which of the selected English words were actually adjectives. If an English word was not identified as an adjective, WordNet was used to find its adjective form by analyzing the derivationally related forms of its lemma. If no such form could be found, then the English word was removed from our list. Both lists were thereafter lemmatized and manually filtered to remove non-evaluative adjectives. This resulted in 2 250 English evaluative adjectives and 1 939 Norwegian evaluative adjectives.

We obtained Russian evaluative adjectives from RuSentiLex (Loukachevitch and Levchik, 2016), which is a list of sentiment-related words and expressions. There are three types of entries in RuSentiLex, depending on their source: 'opinion', 'feeling' and 'fact' (words or expressions that do not express an opinion of the author, but have a positive or negative connotation). Also, each entry is labeled with its part of speech, lemmatized form and polarity, which can be positive, negative, neutral or positive/negative for strong context-dependent semantic orientation. Polysemous words have separate entries for different senses. The current version of the lexicon contains more than 12 thousand words and expressions, which were semi-automatically obtained from existing domain-oriented sentiment vocabularies (initial list), news articles (words with connotations) and Twitter (slang and curse words). For this research we used only one-word adjectives labeled with the 'opinion' source. Since here we do not take into account the differences in the sentiment and polarity of polysemous words, the repeated entries have been removed. In total, there are 2 435 Russian evaluative adjectives in the final dataset.

Table 4.2 summarizes our evaluative adjective lexicons. After acquiring these lists and training word embedding models on the texts created in each decade

[8] https://www.nb.no/sprakbanken/show?serial=sbr-9&lang=nb

under analysis, we were able to move on to the experiments themselves.

4.3 Estimating the speed of semantic change

Our general aim is to measure the degree and strength of temporal semantic shift in evaluative adjectives compared to all other adjective types. This is necessary to assess the intuition that evaluative adjectives are less stable than other words of the same part of speech. Thus, we set out to find evidence across all three languages under analysis.

We would also like to control for frequency and to exclude its influence on the results, since it is known that word frequency often correlates with the speed of semantic change: frequently used words change at slower rates (Hamilton, Leskovec, et al., 2016b).[9] For this reason we experiment both with full sets of evaluative adjectives and with controlled sets limited to one frequency tier.

4.3.1 Methods for quantifying semantic change across time

We measure the speed of semantic change using three methods of comparing the meaning of a word x across two embedding models A and B:

1. Jaccard distance (Jaccard, 1901) between sets of 10 nearest neighbors of x (by cosine distance) in A and B; this is a *local* method.

2. Orthogonal alignment (Hamilton, Leskovec, et al., 2016b): A and B vector spaces are first aligned using the Procrustes transformation, and then cosine distance is calculated between x vectors in two transformed models A_t and B_t; this is a *global* method.

3. Global Anchors (Yin et al., 2018): here, the the intersection of A and B vocabularies ('global anchors', or V_{AB}) is used. The degree of semantic change is defined as the cosine distance between the vector of the cosine similarities of x embedding in A to all words in V_{AB} and the vector of the cosine similarities of x embedding in B to all words in V_{AB}; this is a *global* method.

Note that the 2nd method implies performing Procrustes alignment. We always align two embedding spaces currently under comparison, independent of other time bins. Another possible approach could be to align all five models to one of them, as was done in Kutuzov, Fomin, et al. (2020).

4.3.2 Methods for quantifying change across multiple period pairs

The aforementioned methods measure the distance between the meaning representations of one word in two different embedding models. However, our

[9] Note, however, that this was disputed in Dubossarsky, Weinshall, et al. (2017).

data includes *five* embedding spaces (trained on five consequent decades from the 1960s to the 2000s). In order to quantify the speed of semantic change across the whole time span sequence, we propose two techniques for estimation of semantic change across multiple points:

1. '**Mean distances**': simple mean between the four pairwise distances ('1960s to 1970s', '1970s to 1980s', '1980s to 1990s', and '1990s to 2000s'). It measures the degree of 'semantic jitter' that the word undergoes: it is not necessarily a steady movement into one direction, but can instead consist of fluctuations around one center point or points.

2. '**Mean deltas from the 1960s**': here, at each decade, we calculate the distance δ of the current word representation to its representation in the 1960s (the initial point of our time sequence). If δ has increased, one point is added to the word's score (initialized as 0); if δ has decreased, one point is subtracted. Then, the average score is calculated for each word. The rationale behind this is to measure how steady the shift in meaning is from the initial point for a given word.

   The score here will be low for the words which fluctuate but do not really substantially change their semantics. At the same time, it will be high for consistent cases (like, for example, the English adjective '*solid*' steadily shifting toward denoting not only qualities of materials, but also generally being of good quality). See Figure 4.2 for an example of how a word can first move away from the original meaning, but then start to slowly return back. It shows the trajectory of the Russian adjective 'бескомпромиссный' ('*uncompromising*'). It first moved from the sense of 'CYNICAL, RUTHLESS' closer to a more positive sense of 'PASSIONATE', but then returned back to 'CYNICAL'.

Both 'mean distances' and 'mean deltas from the 1960s' can be used with any method of word meaning comparison, from the three described above. Thus, overall we have 6 scores to assign to each word in our word lists, for all possible combinations of the techniques.

We work with two word lists for each language: the one with *evaluative* adjectives (extracted from sentiment lexicons) and another with what we will refer to as *fillers* or distractors: that is, simply all other adjectives present in the vocabularies of all five models for the current language. We compare the scores for semantic change speed of the words in the first list to those in the second one. If the average values are significantly different with the Welch's T-test p-value not exceeding 0.1[10], we conclude that one type of adjectives is more subject to diachronic semantic change than the other, and report the t-statistics of the difference between the averages. If, on the other hand, the p-value exceeds the

[10] The p-value threshold of 0.1 was used intentionally, instead of the more standard 0.05. We could as well use 0.05, and it wouldn't change the final results of the chapter (the original hypothesis would still be rejected). The reason behind choosing 0.1 was to be able to show that some differences in the speed of semantic change between evaluative adjectives and fillers can be found, but they are rare and fragile even with a very permissive p-value threshold.

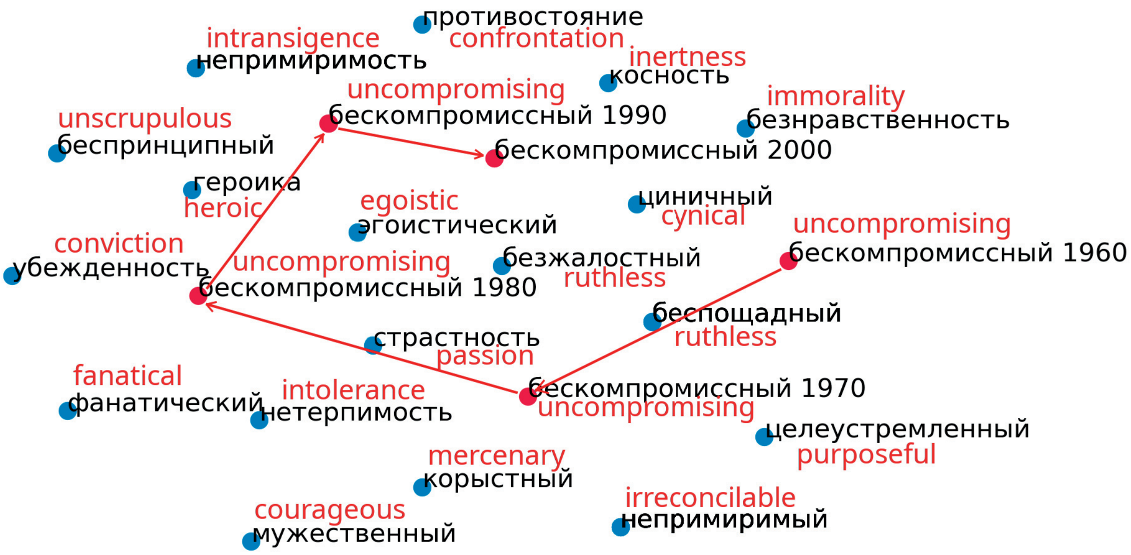


Figure 4.2: Alterations in meaning of the Russian adjective 'бескомпромиссный' ('*uncompromising*'): from 'беспощадный' '*ruthless*' over 'фанатический' '*fanatical*', 'страстность' '*passion*', to 'убежденность' '*conviction*', 'героика' '*heroic*' to 'непримиримость' '*intransigence*', 'противостояние' '*confrontation*'. English equivalents are given in red.

0.1 threshold, we conclude there is no significant difference between two lists, and report the t-statistics as 0.[11] In the next section we provide and discuss the results produced using the aforementioned techniques.

4.4 Experimental results

Table 4.3 presents the results calculated in the way described in the previous section. Positive t-statistic values mean that evaluative adjectives change faster than other types of adjectives, according to particular metrics; negative values mean they change slower; zero values denotes there were no statistically significant differences. We also report the number of filler adjectives ('# fillers') for each language. Recall that 'English10' describes an additional experiment employing exactly the same methods, but with 10 diachronic word embedding models for English, starting from the 1910s, not the 1960s. We report the 'English10' scores in italics, to emphasize that this is an extra experiment, not directly comparable to the main ones.

As can be seen, across all languages, evaluative adjectives fluctuate less (as measured by the 'mean pairwise distances') with all methods, except for Global

[11] Full unabridged tables available at https://github.com/ltgoslo/diachronic_multiling_adjectives/tree/master/full_tables.

	English10	**English**	**Norwegian**	**Russian**
# fillers	*6 746*	8 994	3 989	7 535
Frequency difference	*0.00001*	0.00001	0.00003	0.00001
Method		**Mean pairwise distances**		
Jaccard	*-11.32*	-11.08	-4	-15.05
Procrustes	*-17.34*	-15.52	-5.04	-12.01
Global Anchors	*0*	11.91	-4.40	12.62
		Mean deltas		
Jaccard	*10.67*	3.28	0	0
Procrustes	*8.73*	2.98	0	3.92
Global Anchors	*10.39*	3.57	3.24	3.11

Table 4.3: Differences in the intensity of semantic change between evaluative adjectives and fillers. Positive values correspond to evaluatives changing significantly faster, and vice versa.

Anchors applied to English, English10 and Russian. We will give a possible explanation for this exception in the next subsection.

At the same time, the majority of methods agree that evaluative adjectives are more likely to steady shift in one direction, farther and farther away from the original meaning (as measured by the 'mean deltas from the 1960s'). This is less expressed for Norwegian (with the Jaccard and Global Anchors methods, the difference between the two types of adjectives was not significant).

4.4.1 Experimental results after controlling for frequency

As already mentioned before in this thesis, the speed of semantic change can correlate with word frequencies, although the previous work provides different reports on whether frequent words actually change *faster* or *slower*. The 'Frequency difference' row in Table 4.3 shows the difference between average word frequencies in the evaluative adjectives lists and the fillers lists (expressed as word probabilities relative to corpora sizes). All these values are statistically significant and positive. They show that evaluative adjectives in our dataset are on average more frequent than other adjectives.

Table 4.4 indicates that there are indeed statistically significant correlations between word frequencies and the scores returned by all our methods for measuring the intensity of temporal semantic shift, across all languages. More frequent words consistently get lower scores from the 'mean distances' technique.[12] Vice versa, they get higher scores from the 'mean deltas' technique,

[12] It seems to support the law of conformity from Hamilton, Leskovec, et al. (2016b)

	English10	English	Norwegian	Russian
Method		**Mean distances**		
Jaccard	*-0.38*	-0.37	-0.33	-0.32
Procrustes	*-0.19*	-0.19	-0.21	-0.17
Global Anchors	*0.21*	0.29	-0.08	0.11
		Mean deltas		
Jaccard	*0.09*	0.05	0.10	0.08
Procrustes	*0.12*	0.07	0.12	0.08
Global Anchors	*0.20*	0.07	0.12	0.05

Table 4.4: Correlation of semantic change speed and normalized word frequency across all adjectives (evaluative and fillers). Positive values correspond to frequent words changing significantly faster, and vice versa.

suggesting that frequent words fluctuate less decade-to-decade, but at the same time they are more prone to a slow and steady semantic drift in a particular direction.

An interesting observation can be made about the behavior of the Global Anchors method in Table 4.4. It repeats exactly the phenomenon we already saw in Table 4.3: the English, English10 and Russian values for this method are different in their sign from all other values in the 'mean distances' part. In this case, Global Anchors predictions are positively correlated with frequency: the more frequent the word is, the higher its semantic change score tends to be, which is directly opposite to the behavior of the other two methods.

It seems that all pairwise semantic change estimation techniques are biased by lexical frequencies, but they are biased differently. While Jaccard and Procrustes tend to yield *lower* semantic change scores for frequent words (on average), Global Anchors tend to yield *higher* change scores for the same words. This bias of the Global Anchors is not manifested for the Norwegian dataset (the differences and correlations there are essentially the same as with Jaccard and Procrustes). We believe the reason for this behavior is that our Norwegian dataset has the lowest number of fillers of all three languages. Arguably, this reduces the influence of the low-frequency long tail of fillers for which the Global Anchors yields low change scores. This is also why the Global Anchors returned no significant differences for English10 in Table 4.3: since English10 deals with 10 diachronic embedding models instead of 5, the number of filler adjectives is lower than in the regular English dataset (the intersection of 10 vocabularies is naturally smaller than the intersection of five vocabularies): 6 746 versus 8 994. Again, the omitted fillers were the ones with the lowest frequencies, which led to less overall frequency difference between evaluatives and fillers. Because of

	English10	**English**	**Norwegian**	**Russian**
# fillers	*863*	1 133	571	929
Frequency difference	*0*	0	0	-0.00002
Method		**Mean distances**		
Jaccard	*0*	0	-1.68	-2.54
Procrustes	*-7.33*	-4.77	-3.24	-5.03
Global Anchors	*-6.57*	-3.70	-4.07	0
		Mean deltas		
Jaccard	*3.31*	0	0	-2.44
Procrustes	*0*	0	2.94	0
Global Anchors	*4.95*	0	0	-1.79

Table 4.5: Difference in the intensity of semantic change between evaluative adjectives and fillers (frequency > 100). Positive values correspond to evaluatives changing significantly faster ($p < 0.1$), and vice versa.

that, Global Anchors was 'de-biased' to some extent in English10 (but not as strongly as in Norwegian, with its 3 989 fillers) and did not show any significant semantic change difference.

It is yet to find out what are the underlying reasons for this varied behavior of semantic change estimation methods depending on word frequency. But what is obvious is that different frequencies of evaluative adjectives and fillers introduce undesired noise, and it would be beneficial to get rid of it, so that the experiment is more controlled.

To control for the influence of the frequency factor in comparing evaluative and non-evaluative adjectives, we have to make the average frequencies of both lists more similar. Since we observed that evaluative adjectives are more frequent, we decided to use a frequency cutoff threshold. All adjectives with corpus frequency in at least one decade lesser than the threshold (which is a hyperparameter) were removed from the word lists (both evaluative adjectives and fillers).[13] This allowed us to get rid of the long tail of low-frequency adjectives, and make both lists more similar with regards to frequency in all three datasets. In Table 4.5, we report the results using a threshold of 100; results with the thresholds of 50, 200 and 500 are comparable.

Table 4.5 shows that the number of fillers has naturally declined after introducing the frequency threshold (it is now in the hundreds, not thousands). The number of the evaluative adjectives has also declined: not as strongly, but enough for the datasets to still contain much less evaluative adjectives

[13] We did not down-sample the evaluative adjectives instead, since they are the main focus of this study, and we did not want to reduce their number (which is not huge to begin with).

then fillers. Also, the 'Frequency difference' row indicates that this time we managed to eliminate any statistically significant difference between evaluative and non-evaluative word lists for English and Norwegian. For the Russian data, the situation has even reversed: now evaluative adjectives are on average *less* frequent. Note that word frequencies are not distributed normally, which means that the Welch t-test as a measure of statistical significance can be misleading here. Still, one can see that the absolute differences have changed from being all positive to being of zero or negative value; within the current research, we take it as enough degree of evidence.

The overall results for the 'mean distances' method did not change or even became more expressed. The most important change after the introduction of the frequency threshold is that the Global Anchors is no longer an outlier. It now tells the same story as the other two methods: evaluative adjectives shift is less expressed. This supports our guess that this method's difference from Jaccard and Procrustes was due to the influence of word frequencies. Overall, when controlled for frequency, evaluative adjectives still seem to be *less prone* to 'fluctuating' semantic change. For the 'mean distances' technique, of nine returned scores (three languages by three methods), none reports faster change for evaluative adjectives, with seven reporting slower change, and two reporting no difference at all. Thus, in this respect, evaluative adjectives are more semantically stable than other adjectives. This makes us reject the initial hypothesis about them shifting faster. The additional experiment with the English10 dataset yields exactly the same result.

For the 'mean deltas' technique, filtering out the low-frequency words led to the differences between evaluative and non-evaluative adjectives losing their statistical significance (see zeroes in Table 4.5 cells) with all methods for English. For Norwegian and Russian the results are also similar. Norwegian behaves almost like English (differences not statistically significant except for the Procrustes alignment showing faster change). For Russian, the Procrustes alignment method, vice versa, stopped showing any difference. Instead, the Jaccard and Global Anchors methods turned the opposite (as compared to the experiment without controlling for frequency) and now show that the evaluative adjectives change slower.

Thus, for the 'mean deltas' technique, of nine returned scores, only one still reports faster change for evaluative adjectives, with two reporting slower change, and six reporting no difference at all. We believe this means that the experiments do not support the hypothesis of any specificity of evaluative adjectives with respect to the 'steadiness' of diachronic semantic shifts. No stable differences can be observed here within the time span of five decades.

The picture is more interesting with the English10 dataset, where two methods out of three resulted in evaluative adjectives *more prone* to steady semantic shift in a particular direction. This is not an unanimous vote (and is limited to only language), but still suggests that on longer time spans the behavior of evaluatives can indeed be different from other adjectives in this aspect. Considering the nature of our 'mean deltas' technique, this seems natural: on a longer time span, a slow steady movement can be easier to detect than on a shorter span.

4.5 Limitations

The experiments in this chapter have some limitations, which we describe in this section.

First of all, sentiment lexicons as sources of evaluative adjectives are by all means only proxies. It is quite probable that there are evaluative adjectives beyond sentiment lexicons, and vice versa. In the future, it is possible to refine the datasets and probably come up with more linguistically justified word lists.

Second, five time points (decades in our case) might be not enough to reach convincing conclusions. However, these were the time spans for which we had access to reasonable amounts of reliable textual data for all three languages. It was very important for us that our experiments involved several languages, not English only. Nevertheless, we still reproduce our experiments on 10 decades for English, yielding largely the same results. Overall, historical corpora are often of limited size, and this is one of acknowledged challenges for the lexical semantic change detection field.

Third, better techniques to control for frequency can be devised, not limited to cutting the long tail of low frequency words. It is possible to generate fillers in a more intelligent way: for example, picking one random filler for each evaluative adjective, mimicking its corpus frequency (recall that we simply used all the remaining adjectives instead).

Finally, although we used well-known methods of lexical semantic change estimation across word embedding models (many of them were described in Chapter 2 of this thesis and evaluated in the previous work), there is still a need to further evaluate the methods themselves.

One option here it to use the SentProp historical sentiment dataset from Hamilton, Clark, et al. (2016). It describes the variation of English word sentiment over historical time-periods, and contains about 2000 adjectives per each decade from the 1960s to the 2000s, annotated with mean sentiment.

As a sort of evaluation, we calculated the adjective-wise differences in the SentProp sentiment scores between each consecutive decade. Then we found the correlations between these differences and the degrees of semantic change returned by our methods described above.The aim here was to find out which of our algorithms produces results better correlated with the output of another system. Interestingly, we did not find statistically significant correlation with the SentProp for any of the employed algorithms. This is not critically wrong, since the SentProp is not human-annotated data: it was also created automatically. Still, this observed discrepancy is interesting and we plan to research it further in the future.

One of the reasons for it can be that it is in general difficult for distributional representations to handle the differences between *antonyms* (Ono et al., 2015; Z. Chen et al., 2015). At the same time, antonymic changes constitute a significant part of diachronic shifts in SentProp. It is still an open question whether reliable antonym treatment is possible at all with models based on distributional signals from corpora. There is an ample room for further research here.

4.6 Summary

In this chapter, we measured the intensity of diachronic semantic change in adjectives across three languages (English, Norwegian and Russian) and five decades (1960s, 1970s, 1980s, 1990s, 2000s), to test whether evaluative adjectives change faster or more intensely than other adjectives. We did not propose any new models here, but tested the applicability of the existing ones to a concrete linguistically motivated problem.

Our results show that, contradictory to the initial hypothesis, evaluative adjectives change over time *less intensely* (statistically significant at $p < 0.1$), if we measure change as the mean of pairwise differences between successive decades, and not as a steady drift in one particular direction. At the same time, when measuring the probability of steadily 'shifting' from an original meaning across time, evaluative adjectives *do not differ from other adjectives at all* (on any statistically significant level).

These observations are not frequency artifacts, since we observe the same behavior when controlling for word frequencies. These controlled experiments additionally allowed us to trace how semantic change detection methods are influenced by frequency in different ways. In particular, it seems that Jaccard distances between the nearest neighbors and cosine distances between Procrustes-aligned models tend to yield *lower* semantic change scores for frequent words, while the Global Anchors method tend to yield *higher* change scores for frequent words.

We also conducted an extra experiment with the increased 'observation window' of 10 decades for English (starting from the 1910s). In this case, two of the three our methods reported more expressed steady drift in one particular direction for evaluative adjectives (but still less expressed for the pairwise differences between successive decades). Our interpretation is that there is no difference between evaluatives and other adjectives in their short-term fluctuations (independent of the width of the observation window, be it five decades or 10). But if we observe language data for a longer time, diachronic embedding-based methods may start to capture a show and consistent movement of evaluative adjectives away from their original meaning. We hope to study this in the future with more languages and more varied observation windows.

To sum up, it seems that evaluative adjectives are not more prone to semantic shifts than other adjectives: at least with the observation window of five decades. Vice versa, with regards to decade-to-decade pairwise shifts, they are even more stable than their counterparts; this holds across different languages and semantic change detection methods.

Diachronic embedding models, word lists and code used at the experiments in this chapter are publicly available, see Chapter 7 for the links.

In this chapter, we took the existing semantic change detection algorithms for granted, assuming that they are fit for the task. But we also need a robust way to evaluate the performance of algorithms which extract information about semantic change from diachronic embeddings. Before moving on to the specially designed diachronic semantic change test sets (most of them published very recently) in Chapter 6, we will first use databases which record armed conflicts occurring in the world. They can be employed as proxies to language changes. A particularly important example of such databases is the Uppsala Conflict Data Program dataset. The next Chapter 5 describes this dataset. It continues with the presentation of how we employed it to evaluate diachronic word embedding-based algorithms with regards to their ability to predict real-world events unfolding in time. This will demonstrate the versatility of semantic change related information which can be captured by distributional vector models.

Chapter 5

Semantic change and world events: armed conflict dynamics

In the previous Chapter 4, we took the existing semantic change detection algorithms for granted and applied them to produce answers for our linguistic question. However, these algorithms should be evaluated as well. As discussed in Chapter 3, proper test sets for diachronic semantic change have only recently started to appear. The majority of world languages still lack such test sets. In this chapter, we attempt to overcome this problem by using language-agnostic historical event datasets to evaluate and probe semantic change modeling methods based on word embeddings. In particular, we focus on armed conflict datasets, containing location and armed group names and the temporal data (when the conflicts started or ended).

Note that semantic change cases we deal with in this chapter are mostly of referential or 'world knowledge' nature, and fall into the context variance span on the semantic proximity gradient (see the Introduction and Chapter 2). We claim that such changes are still semantic, although they are different from semantic shifts proper (acquiring or losing a lexicographic sense).

Armed conflicts manifest well-defined temporally limited real world events: they naturally possess starting and ending dates. They also obviously influence human-generated texts, receiving wide coverage in the news. If we possess the ground truth about when conflicts started and ended, we are then able to evaluate our approaches towards semantic change detection with diachronic word embeddings. Armed conflict data here functions as *distant supervision* (Fang and Cohn, 2016), allowing one to indirectly check the usefulness of the machine learning system when one lacks annotated data. At the same time, armed conflicts tracing through NLP text analysis algorithms can be practically useful for peace research, social studies, information retrieval and other disciplines.

Importantly, this paves the way to evaluating semantic change detection methods for multiple languages. Armed conflicts arguably get approximately the same amount of coverage in the news texts, independent of language.[1] This means it is sufficient to translate the named entities in the dataset to apply it to another language (given enough news texts in this language is available).

Note that there is also a large field within NLP called 'event extraction' (see for example Ji and Grishman (2008) and Hürriyetoğlu et al. (2020)), with its own datasets. However, this is out of scope for this chapter: we do not deal with extracting particular events from particular documents. For example,

[1] This is of course not entirely true: conflicts can be covered in different ways because of state propaganda, or even be intentionally silenced. But this is not specifically related to languages.

analyzing a particular news text to find out whether it contains a description of an armed conflict, and if yes, what are the properties of this conflict, is out of scope. Instead, we are interested in inferring event facts from large corpora as a whole, without mapping events to certain documents, or text chunks within these documents.

Importantly, inferring the facts about armed conflict dynamics is not an aim in itself for us. The temporal event detection task in this chapter is only a proxy to study and evaluate information captured by diachronic distributional representations. Because of that, we do not compare against methods from the event detection field. As already discussed before, what we are doing resembles *probing* of machine learning systems. For example, nobody expects that raw representations at the n^{th} layer of a deep neural architecture will alone outperform the state-of-the-art in syntactic parsing. However, trying to use these representations to solve the syntactic parsing task can help researchers to better understand what the model has learned about language structures (Hewitt and Manning, 2019). In the same vein, by probing diachronic word embeddings for their ability to detect or predict changes in the real world, we can better understand what these models 'know' about the accompanying changes in lexical semantics. Thus, our aim is not to develop a state-of-the-art event detection system, although admittedly the workflow can look similar, and some of the methods we propose can in principle be used for this task.

5.1 Armed conflict research data

In this section, we describe the Uppsala Conflict Data Program, which stores and processes information about armed conflicts happening in the world. Quoting their web page:[2]

'*The Uppsala Conflict Data Program (UCDP) is the world's main provider of data on organized violence and the oldest ongoing data collection project for civil war, with a history of almost 40 years. Its definition of armed conflict has become the global standard of how conflicts are systematically defined and studied. UCDP produces high-quality data, which are systematically collected, have global coverage, are comparable across cases and countries, and have long time series which are updated annually. Furthermore, the program is a unique source of information for practitioners and policymakers.*'

The UCDP Conflict Encyclopedia provides ready-made datasets featuring regularly updated information about armed conflicts. There exist other similar databases, like, for example, those used in Zukov Gregoric et al. (2016) (casualties in Iraq war) or in Mueller and Rauh (2017). However, they are either too small or don't have enough coverage for our purposes, or are not publicly available. Thus, in the rest of this chapter we stick to the UCDP datasets.

In the following subsections, we briefly outline the history of armed conflict research and peace studies. We then move on to the UCDP project itself and its datasets. After describing their nature and the process of their creation, we

[2] https://www.pcr.uu.se/research/ucdp/about-ucdp/

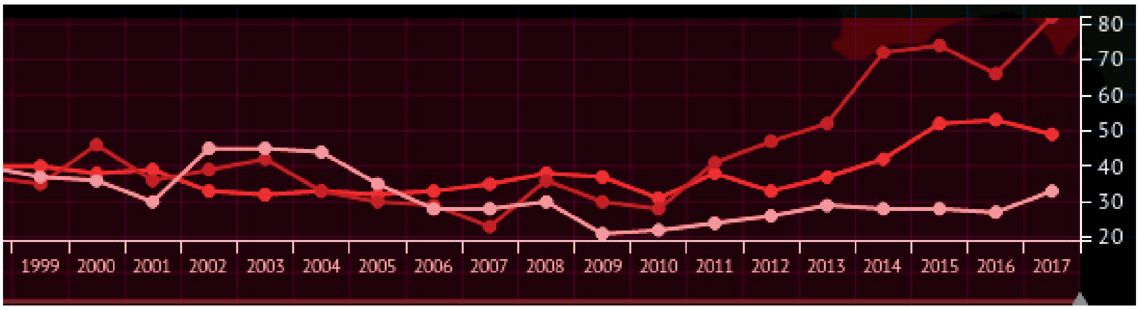


Figure 5.1: Number of armed conflicts in the world per year. Colors denote conflict types: pink stands for one-sided violence, scarlet red stands for state-based violence, and dark red stands for non-state violence (source: Uppsala Conflict Data Program).

show how they can serve as a distant supervision signal to evaluate semantic change detection methods. Additionally, finding out what information about real-word events is captured by diachronic word embeddings is an interesting research aim in itself.

5.1.1 Conflict Research Overview

Conflict research is a well-established academic discipline within social studies and is increasingly relevant today. It can be seen as a branch of peace research: peace is absence of conflict, and conflict research studies the military conflicts which can jeopardize peace: their origins, dynamics and resolution (Wallensteen, 2013).

The importance of such studies is obvious: hundreds of armed conflicts arise globally each year, and millions of human beings are affected by them. Figure 5.1 (taken from the UCDP website) reflects the dynamics of the amount of different kinds of armed conflicts in the world in the recent years. It shows that the number of armed conflicts is unfortunately growing (although the number of casualties is steadily decreasing), which means that conflict research becomes even more topical. Efficient ways to reduce the number of armed conflicts and to build peace can be devised only if science can explain how these conflicts arise, what factors influence the possibility of conflicts, and how exactly they end, including how different peace agreements work.

5.1.1.1 History of conflict research

Proper academic conflict research started from the seminal work of Wright (1942). It quickly evolved into a full-scale academic discipline, studying both direct violence in the form of armed conflicts and so called structured (indirect) violence in the form of lessening human life span in other ways apart from armed conflicts.

In this chapter, we deal only with the former type of violence as it is much easier to quantify into events with specific start and end dates.

Both qualitative and quantitative studies must be supported by properly collected data. This was the reason for the appearance of conflict data collection initiatives, pioneered by the 'Correlates of War' (COW) project (Singer and Small, 1972). In the eighties it was followed by the Uppsala Conflict Data Program (UCDP) in Uppsala University, Sweden (Gleditsch et al., 2002). It features the online *UCDP Conflict Encyclopedia* with detailed descriptions of the conflicts and armed groups participating in them, and rich metadata. Its collaboration with the Peace Research Institute in Oslo (PRIO), Norway allowed for augmenting this with data from the year 1946 for some selected metadata fields (UCDP/PRIO dataset). We describe the UCDP datasets in more details in section 5.1.2.

5.1.1.2 General usage of the conflict datasets

Systematically compiled and manually annotated conflict datasets are in high demand in today's world. Eck (2005) enumerates three main types of conflict data users:

1. *Policy makers* who need the data to make informed decisions especially related to peacemaking. For them, it is important that the datasets are regularly updated and reflect the current situation correctly.

2. *Academics doing historic research.* They are mostly interested in qualitative data like conflict summaries to draw on them as the basis for further inquiries and interpretations.

3. *Academics doing quantitative research* on a large scale. They possibly benefit most from the conflict datasets, as quantitative conflict studies are hardly imaginable without big conflict data, and datasets like UCDP and others are the only source of this information. The author of the present thesis clearly belongs to this last group.

All of these groups are interested in scrupulously and systematically curated datasets. All of them arguably can benefit from using the methods of semi-automated armed conflict tracing from news texts proposed in this chapter.

5.1.2 Uppsala Conflict Data Program

The UCDP/PRIO Armed Conflict Dataset[3] maintained by the Uppsala Conflict Data Program[4] and the Peace Research Institute Oslo (PRIO)[5] is a geographical and temporal dataset with information on armed conflicts, both internal (within one national state) and external (crossing state borders), where at least one

[3] https://www.ucdp.uu.se/

[4] http://www.pcr.uu.se/research/ucdp/program_overview/about_ucdp/

[5] https://www.prio.org/Data/Armed-Conflict

party is the government of a state, in the time period from 1946 to the present (Gleditsch et al., 2002). The Armed Conflict Dataset is updated on a constant basis and is primarily intended for academic use in statistical and macro-level research. The collection of the dataset has started in the mid-1980s under the name 'Conflict Data Project', but since then evolved constantly. In the autumn of 2003 the amount of work on conflict data collection led to a change in the name of the project and it was thus turned into the 'Uppsala Conflict Data Program'.

Two notions are essential for the UCDP datasets: *event* and *armed conflict*. Basically, an *event* is an incident where armed force was used by an organized actor against another organized actor, or against civilians, resulting in at least one direct death at a specific location and a specific date (see more on this in Sundberg and Melander (2013)). Note that *armed force* here means the use of arms in order to promote the parties' general position in the conflict, resulting in deaths (in turn, *arms* means any material means, e.g. manufactured weapons but also sticks, stones, fire, water etc.). *Organized actor* can be a government of an independent state, a formally organized group or an informally organized group according to UCDP criteria (Sundberg and Melander, 2013).

Such *events* can evolve into full-scale *conflicts*, defined as contested incompatibilities that concern government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths. Note that it does not need to be a single event resulting in 25 deaths: this number can be accumulated over several events over time.

The UCDP datasets constitute a case of quantitatively encoded social data spanning over decades. Below we describe the datasets in more detail and briefly overview how we use them.

5.1.3 UCDP Georeferenced Event dataset

The UCDP yearly releases several public datasets with data on major conflicts.[6] One of the most important of those is the Georeferenced Event dataset (hereafter GED). The current version of GED at the time of writing is 19.1, released in 2019.

GED lists and describes armed conflict events themselves (Croicu and Sundberg, 2015). It includes three types of organized violence:

1. state-based conflict (involving at least one government actor);
2. non-state conflict (violence between two non-governmental actors);
3. one-sided violence (towards civilians).

Thus, the entities in this dataset are armed conflict events. The 19.1 version of GED contains 152 616 events and about 1 500 different conflict actors. Conflict actors can be either governments or rebels, separatists, other insurgent groups,

[6] https://www.ucdp.uu.se/downloads/

etc. Of all the events in the dataset, 28 929 happened in Afghanistan, 15 328 in India, and 7 473 in Iraq. These three countries contribute to more than third of all armed conflict events.[7] GED is a rich source of data about exact temporal moments when armed conflicts within particular locations started or ended.

The UCDP creates GED by examining news texts. The textual data comes initially from global newswires and BBC Monitoring data sourced from the Dow Jones Factiva aggregator (sometimes other secondary sources are also used). UCDP extracts and processes only those texts which contain the following token-based search patterns:

1. kill*
2. die*
3. injur*
4. dead
5. death*
6. wounded
7. massacre*

The texts which do not contain these patters are ignored altogether. Note that this means that the datasets are inherently biased towards a particular kind of armed conflict descriptions in the text. This can in theory influence the results of any research based on the UCDP data. However, the extent and the importance of this hypothetical bias is unknown, and in the further text we ignore it.

The extracted documents are then manually checked by the UCDP human annotators. Each document is annotated with metadata by at least two human experts, all the controversial decisions are discussed and resolved in a reconciliation procedure. If a document is found to be irrelevant, it is marked as ‘Negative’ (not describing an armed conflict) and further ignored. If the annotators consider the document as describing an armed conflict event, it ends up as ‘Positive’, and eventually is linked to some of the events in the GED dataset.

It is important that the GED dataset extensively describes armed conflicts in many details (see the Appendix A). It contains not only dates but also the meta information related to armed conflicts: conflict type, who are the actors, what are the casualties, etc. This provides us a great amount of temporally changing gold data which we can try to extract from diachronic word embeddings. This can also help to evaluate some aspects of semantic change detection approaches,

[7] Note that at the time being the dataset does not include data for Syria. The maintainers claim the final product is not releasable at this time with the same level of consistency and clarity as other UCDP GED data.

since real-world events often result in cultural (or 'event-based') semantic drift. We describe our evaluation mode in more details in subsection 5.1.4.

Note that although the UCDP datasets are structured data (in addition to the news texts themselves), they still require at least some pre-processing before they can be used for evaluation of our approaches. We inferred all the necessary information from the fields of metadata. The names of the entities (countries and actors) had to be normalized so that every entity is mapped to exactly one lexical unit. The UCDP Actor List dataset was of use for this task; it contains many variants of spelling the names of entities. Other possible resources to help normalization are JRC Names[8] and Wikipedia.

5.1.4 Evaluation of semantic change detection with the UCDP datasets

The abundance of data fields in the GED dataset makes it possible to apply it to many tasks. In this work, in particular, we use the dataset to study and evaluate information captured by diachronic distributional representations. We do it mainly by referring to the start and end dates of armed conflicts.

As stated in Chapter 1, one of the aims of this thesis is to study robust and reliable ways to extract semantic change related information from the analysis of diachronic word embeddings. Fortunately, the existence of datasets like GED also provides us with much needed evaluation data. Particularly, it is a source of ground truth about the starting and ending points of armed conflicts and their development. When accompanied with news corpora like Gigaword (Parker et al., 2011), SignalMedia (Corney et al., 2016), or News on Web (NOW)[9], it allows us to evaluate how good our approaches are for extracting cultural semantic shifts from raw texts. Besides significant academic interest, this topic has a practical application: we are going to implement and test systems which would use traced semantic change (based on the embeddings trained on news texts) to predict starting or stopping armed conflicts.

Let us recall our general workflow. We train diachronic distributional word embedding models (making them comparable through incremental training or through some variation of alignment) on the news texts, using the sequences of time spans determined by the periods for which we have data. For example, one can train $12 * 3 = 36$ models for three years, each time adding texts produced in the next month. This way, we receive a sequence of embedding models reflecting different time spans[10]. We can compare them against each other to discover change in the representations for particular entities. In this case, we start with geographical locations as entities. We extract the data related to events in these locations from the GED dataset: we are particularly interested in the starting of armed conflicts in the previously peaceful locations and in the termination of such conflicts in the locations which were at war before. In other words, we look

[8] https://data.europa.eu/euodp/en/data/dataset/jrc-emm-jrc-names

[9] https://corpus.byu.edu/now/

[10] See more on that in Chapter 3

for 'breaking points', where the social and political landscape change, which can lead to significant change in the news texts covering a particular location.

Then, we analyze the output of our approaches to modeling semantic change based on distributional word embeddings and trace diachronic changes in semantic representations for the entities mentioned above. We hypothesize that when an armed conflict starts or stops in some location, it is reflected in the texts mentioning this location. This, in turn, provokes specific changes in the embeddings of lexical entities denoting the area or the active armed groups in this area. Most often, these changes belong to the context variance type we already mentioned in the Introduction and in Chapter 3: the entity begins to be employed in entirely new contexts and it now gives rise to different connotations in the reader's or listener's mind. Ideally, these changes should correlate with the ground truth on armed conflicts extracted from the GED dataset. The correspondence of predicted results to this ground truth provides the basis for our evaluation metrics.

5.1.5 Constructing a gold standard dataset

As stated before, to properly evaluate our approaches towards semantic shift detection, we need precise data on armed conflicts starting and ending. For this, one can employ the subset of the GED dataset called UCDP Conflict Termination dataset,[11] containing entries on starting and ending dates of about 2 000 conflicts up to year 2015. Another possibility is to extract data directly from the GED (it provides more unnecessary data fields, but at the same time contains more recent information).

We omit the conflicts where both sides were governments (about 2% of the entries), for example, the 1998 conflict between India and Pakistan in Kashmir. The reason for this is that with these entries, static distributional models have a hard time telling the name of the state (conflict actor) from the name of the territory (conflict location): '*Iraq*' can mean both the 'GOVERNMENT OF IRAQ' and the 'IRAQ AS A LAND AREA'. In principle, this is just a technical issue and can arguably be solved by, for example, using contextualized embedding architectures (see the next Chapter 6, where we use them for semantic change detection). However, such architectures were not yet available at the time of working on this particular part of the thesis. Additionally, the 'government – insurgent group' opposition is more pronounced and asymmetric, and serves better as an example of a semantic relationship for the purposes of section 5.3 below. Thus, we analyze only the conflicts between a government and an insurgent armed group of some kind (these conflicts constitute the majority of the UCDP data anyway).

Another group of conflicts we omit in the experiments described further in this chapter is where at least one of the sides has a corpus frequency of less than 100 in the corresponding text collection. The rationale for this decision was that these conflicts have too little contextual coverage in the corpus for

[11] https://ucdp.uu.se/downloads/monadterm/ucdp-term-conf-2015.xlsx

distributional models to learn meaningful representations for them. These cases usually constitute about 1% of the entries, depending on the particular corpus (Gigaword, NOW, etc.).

The first version of the resulting test set based on the Gigaword corpus (Parker et al., 2011) and the UCDP Conflict Termination dataset covers the time span from 1994 to 2010 (the last year of Gigaword). It mentions 52 unique locations (with '*India*' being the most ubiquitous), 673 unique armed conflicts, and 128 unique armed insurgent groups (with '*ULFA*' or '*United Liberation Front of Assam*' being the most ubiquitous). Location names have an average per-year corpus frequency of 17 749, and for the armed groups this value is 570 (recall that the average yearly corpus size of Gigaword is about 300 million word tokens). Naturally, in both cases the frequency distribution follows the power law (several high frequency items with many low frequency items, the standard deviation much higher than the average value), but still we can see that the armed group names are much less frequent than the location names, which is expected.

The UCDP dataset also includes the intensity level of the conflict in each particular year: 493 conflicts are tagged with the intensity level 1 (between 25 and 999 battle-related deaths), and 180 conflicts with the intensity level 2 (at least 1 000 battle-related deaths). For location–year pairs with no records in the UCDP dataset we assign the tag 0, indicating that there were no armed conflicts in this location at that time.

The resulting test set which we dub Armed Conflicts Evaluation Test Set is available at https://github.com/ltgoslo/diachronic_armed_conflicts. Further on, we also produced several other versions of this test set, see the resource list in Chapter 7 and the sections below.

5.1.6 Summary

The UCDP armed conflict data allowed us to compile a gold standard data set containing the start and end dates of armed conflicts throughout the world across several decades. In the following sections, we will extensively use this test set and its derivatives to study how this information is captured by diachronic word embeddings via shifts in the context variance of the corresponding entities. In particular, we will describe two experiments employing the Armed Conflicts Evaluation Test Set. The first one tries to detect armed conflicts on a year-to-year basis for a given location (country). The second one traces how semantic relations between locations and violent armed groups (insurgents) change with the time.

5.2 From nearest neighbors to anchor words: tracing armed conflicts

Now that we have the gold standard, it is possible to evaluate our approaches against it. In this section, we trace changes in the local semantic neighborhoods of country names (as measured by static embedding-based methods), applying it to the downstream task of predicting changes in the state of conflict for 52 countries at the year-level. The input data here is the Armed Conflicts Evaluation Test Set, and a corpus of news texts published in the years from 1994 to 2010, organized in sequential pairs of years.[12]

The results of this experiment provide insights about the performance of several semantic change detection techniques in capturing real-world events: 1) cosine similarities between Procrustes-aligned vector spaces, 2) the domain-specific anchor words method proposed by us and similar to the 'local anchors' method from Zhang et al. (2016). We show that using domain-specific anchor words outperforms the Procrustes/cosine method when applied to the armed conflict data (the selection of anchor words is described below). However, the issue of choosing the method of quantifying differences between representations of one and the same word in two embedding spaces is more complicated.

5.2.1 Data and labels encoding

We cast armed conflicts state detection from news texts as a classification task with three classes:

1. Nothing has changed in the country conflict state year-to-year: either the country remained peaceful, or a conflict continued ('stable' class);
2. Armed conflicts have escalated in the country year-to-year ('war' class);
3. Armed conflicts have declined in the country year-to-year ('peace' class).

We represented the Armed Conflicts Evaluation Test Set as a set of data points equal to the differences (δ) between the location's conflict state in the current year and in the previous year. Conflict state is equal to 0 if no armed conflict was observed in a particular country and year. If the UCDP records an armed conflict for this time and location, the conflict state value equals to the conflict intensity level. It can be either 1 or 2: in the Armed Conflicts Evaluation Test Set, 493 conflicts are tagged with the intensity level 1 (between 25 and 999 battle-related deaths), and 180 conflicts with the intensity level 2 (at least 1 000 battle-related deaths). If there were several conflicts in the location in a particular year, we used the average of their intensities. As a result, we have 832 data points in total (52 locations × 16 year pairs).

As an example, in Congo, the conflict with the intensity level 1 had terminated when transitioning from 2001 to 2002. Thus, for the data point 'congo_2002', $\delta = 0 - 1 = -1$. 0 here means there was no armed conflict in Congo in 2002, and 1 means there was an armed conflict of the intensity 1 in Congo in 2001. After that, there were no changes (each new δ had the value of 0) until 2006, when armed conflicts resumed again with the intensity of 1. Thus, for the 'congo_2006' data point, $\delta = 1 - 0 = 1$.

[12] Parts of this section were previously published as Kutuzov, Velldal, et al. (2017b).

However, for practical reasons it is more useful to predict a human-interpretable class of the conflict state change, rather than a scalar value. A version of this test set was produced where δ values were transformed to classes, following the equation 5.1 (other thresholds are of course possible as well).

$$class = \begin{cases} changing - to - war & \text{if } \delta \geq 0.5 \\ changing - to - peace & \text{if } \delta \leq -0.5 \\ stable - state & \text{otherwise} \end{cases} \tag{5.1}$$

The 'change' classes *changing-to-war* and *changing-to-peace* constitute 10% and 11% of the data points respectively. Thus, they are minority classes and we are mostly interested in how good the evaluated methods are in predicting them, not the majority *stable-state* class. Also, from the practical point of view, the changing points are certainly more interesting. Below we describe the evaluated approaches.

It is important to note that in Hamilton, Leskovec, et al. (2016b) and other previous work on semantic change modeling, proper names were mostly filtered out: their authors were interested in global semantic shifts for common nouns. In contrast to this, we here make proper names (countries and other named entities) our main target. In this chapter, we are mostly interested in what is happening to the referential meaning of this or that named entity, not in whether there were changes in the denotational meaning of a common noun. Thus, this setup is similar to tracing drift in world knowledge associated with a word in language. As already discussed in Chapter 2, we follow the opinion of Geeraerts (1997) that there is no clear borderline between the 'linguistic meaning' and 'world knowledge'. Thus, these shifts (some country starts being associated with war), from our point of view, do belong to the domain of lexical semantic change, although there is no observed change in lexicographic senses.

5.2.2 Detecting changes in armed conflict states

As we are dealing with temporal data, we experiment with different methods for extracting chronological information from word embedding models. In this subsection, these methods are described, along with different types of diachronic word embeddings and domain-specific anchor words we employed (see below).

As a source of the training data, we used the Gigaword English news corpus (Parker et al., 2011), containing about 5 billion words. All Gigaword texts are annotated with publishing date, so it is trivial to compile yearly corpora starting from 1994 up to 2010. Then, we trained three sets of 17 yearly word embedding models, differing in the way they represent yearly time bins:

1. models trained from scratch on the corpora containing news texts from a particular year only (dubbed *scratch* hereafter);
2. models trained from scratch on the corpora from the particular year and all the previous years (dubbed *cumulative* hereafter);

3. incrementally trained models (dubbed *incremental* hereafter).

The last type was most interesting for us: here one and the same vector space is incrementally updated with new data. However, unlike the original suggestion by Kim et al. (2014), we also update the model's *vocabulary* as new data arrive.[13] Our hypothesis was that this can help coping with the inherently stochastic nature of predictive distributional models. However, this turned out to be not entirely true in this case (see below).

We used the Gensim library (Řehůřek and Sojka, 2010) for training Continuous Bag of Words word embedding models (Mikolov, K. Chen, et al., 2013) on the Gigaword corpus. In terms of corpus pre-processing we performed lemmatization, PoS-tagging and name entity recognition using the Stanford CoreNLP library (Manning et al., 2014). Named entities were concatenated to one token (for example, '*United States*' became '*United::States_PROPN*').

Once the sets of models are there, one can detect semantic change for a given target word wq (in our case, always a location name) and a given pair of models. We compare our *domain specific anchor words* approach to the alignment with orthogonal Procrustes method heavily used in previous work and described before in Chapter 3. The two methods can be described as follows:

1. *Procrustes*: align two models (trained on the current and previous year, M_{cur} and M_{prev}) using the orthogonal Procrustes transformation (Gower, Dijksterhuis, et al., 2004). Then measure cosine distance between the wq_{cur} and wq_{prev} vectors, as proposed in Hamilton, Leskovec, et al. (2016b). Higher distance means stronger change in meaning. This is an example of a *global* approach to semantic change detection.

2. *Domain specific anchor words*: define a set of words related to the semantic categories we are interested in. These words are selected manually, and are called 'anchors', since they serve as measures of the degree of word drift in the vector space. Then, measure this drift of wq towards or away from these anchors' vectors in M_{cur} compared against M_{prev} (in terms of cosine distances). Higher values of drift mean stronger change in meaning.

   This is, instead, an example of a *local* approach to semantic change detection, initially proposed in Zhang et al. (2015) and Zhang et al. (2016) (called 'reference points' there). However, we were the first to enrich this method with using not just the nearest neighbors but domain-specific words in general in Kutuzov, Velldal, et al. (2017b) (also, the task was different). Later, Garg et al. (2018) used a similar approach to analyze the dynamics of ethnic and gender biases in word embedding models, and Yin et al. (2018) developed the same idea further to include all words from the model vocabulary as anchors, thus making it a global approach (we employed this Global Anchors method in Chapter 4).

[13] The model M_{t+1} is initialized with the weights from the model M_t; if there are new words in the $t + 1$ corpus which exceed the frequency threshold, then before the start of M_{t+1} training they are added to its vocabulary and their vectors are initialized with random weights.

The first (Procrustes) method outputs one value of cosine distance for each data point, representing the degree of semantic change, but not its direction. In contrast, the domain-specific anchor words method can potentially provide information about the exact direction of the change. This, in turn, can be quantified in two ways:

1. Quantification mode 1: for each anchor word, calculate its cosine similarity against wq in M_{cur} and M_{prev} (dubbed 'Sim' hereafter);

2. Quantification mode 2: the same, but instead of using the cosine, find the *position* of each anchor in the models' vocabulary sorted by similarity to wq; we normalize by the size of the vocabulary so that rank 1 means the anchor is the most similar word to wq while rank 0 means it is the least similar (we dub this approach 'Rank').

Both quantification methods produce two vectors $\vec{\boldsymbol{r_{prev}}}$ and $\vec{\boldsymbol{r_{cur}}}$, corresponding to the models M_{prev} and M_{cur}. Their dimensionality is equal to the number of the anchor words, and each component of these vectors represents the similarity of wq to a particular anchor word in a particular time period.

To compute the difference between $\vec{\boldsymbol{r_{prev}}}$ and $\vec{\boldsymbol{r_{cur}}}$ (and thus, the degree of a shift), one can either:

1. calculate the cosine distance between these 'second-order vectors', as described in Hamilton, Leskovec, et al. (2016a); we dub this 'SimDist' or 'RankDist', depending on whether 'Sim' or 'Rank' quantification mode was used;

2. element-wise subtract $\vec{\boldsymbol{r_{prev}}}$ from $\vec{\boldsymbol{r_{cur}}}$ to understand whether wq drifted towards or away from the anchors; we dub this 'SimSub' or 'RankSub', again depending on the quantification mode.

In the first case, the output is again one scalar value, and in the second case it is the vector of diachronic differences, with the dimensionality equal to the number of the anchor words. Either of these feature sets can then be fed into any classifier algorithm. To predict the actual direction of the change (or the absence of change), one needs to perform classification into three classes: *changing-to-war*, *changing-to-peace* and *stable-state*.

5.2.2.1 Anchor word sets

To evaluate the approaches described above, we needed a set of anchor words strongly related to the topic of armed conflicts. For this, we simply adopted the list of seven search strings used within the UCDP to filter the news texts for subsequent manual coding, already mentioned above:

1. '*kill*',

2. '*die*',

3. ‘*injury*’,
4. ‘*dead*’,
5. ‘*death*’,
6. ‘*wound*’,
7. ‘*massacre*’.

Additionally, an expanded version of this set was created, where every original anchor word was accompanied with its five nearest neighbors (belonging to the same part of speech) in the word embedding model we trained on the full Gigaword corpus. This resulted in the following set of 26 words (Gigaword corpus frequencies for each word are given in parentheses):

1. ‘*kill*’ (2 444 435)
2. ‘*death*’ (1 318 501)
3. ‘*die*’ (1 148 940)
4. ‘*dead*’ (624 942)
5. ‘*injury*’ (556 580)
6. ‘*injure*’ (513 973)
7. ‘*wound*’ (455 361)
8. ‘*murder*’ (394 336)
9. ‘*killing*’ (304 781)
10. ‘*massacre*’ (97 538)
11. ‘*genocide*’ (96 074)
12. ‘*hospitalize*’ (63 192)
13. ‘*atrocity*’ (53 801)
14. ‘*slaying*’ (49 073)
15. ‘*slay*’ (46 248)
16. ‘*fatality*’ (35 518)
17. ‘*slaughter*’ (34 162)
18. ‘*gun*’ (32 562)
19. ‘*missing*’ (22 773)

20. ‘*perish*’ (18 894)
21. ‘*concussion*’ (18 098)
22. ‘*unaccounted*’ (10 845)
23. ‘*sprain*’ (5 269)
24. ‘*bullet-riddled*’ (3 457)
25. ‘*drowning*’ (3 258)
26. ‘*contusion*’ (1 209)

One can also try to filter the anchor word list by removing low-frequency words, as suggested in Zhang et al. (2015). We did not do this explicitly, since (as seen from the frequencies in the list above) our first set of seven conflict words from the UCDP filtering workflow already contains only high frequency words. All its seven entries can be found in top 10 most frequent words from the second set. Thus, the first set also serves as a testing ground for using only high frequency lexemes.

Ways can be devised to select domain-specific anchor words fully automatically, without using any external seed words. One of the simplest ways to do that (among many others) is to calculate a set of the nearest neighbors of the domain name in a relevant word embedding model. Using the embedding model trained on the full Gigaword corpus and the word ‘*conflict*’ as the domain name, we came up with the following set containing the word itself and its 10 nearest neighbors (evaluated below as ‘automatic anchors’):

1. ‘*conflict*’
2. ‘*strife*’
3. ‘*bloodshed*’
4. ‘*war*’
5. ‘*hostility*’
6. ‘*confrontation*’
7. ‘*bloodsh*’ (incorrect lemmatization of ‘*bloodshed*’)
8. ‘*bloodlet*’
9. ‘*dispute*’
10. ‘*fighting*’
11. ‘*violence*’

Finally, it is also possible to compile a list of anchor words associated with peace instead of conflicts. They would have to be chosen more arbitrarily. Just to serve as an example, we compiled the following list of 'peace anchors' (evaluated below along with the conflict anchors):

1. '*cease-fire*'
2. '*ceasefire*'
3. '*cessation*'
4. '*peace*'
5. '*peacemaking*'
6. '*truce*'

The classification itself was done using a one-vs-rest Support Vector Machine (SVM) classifier (Boser et al., 1992) with balanced class weights (the weight of each class being inversely proportional to class instances frequency in the training data). The features used were either the cosine distance between $\boldsymbol{r_{\vec{prev}}}$ and $\boldsymbol{r_{\vec{cur}}}$ (in the case of 'SimDist' and 'RankDist') or the result of $\boldsymbol{r_{\vec{cur}}} - \boldsymbol{r_{\vec{prev}}}$ subtraction (in the case of 'SimSub' and 'RankSub'). In the first case, we have only one feature, while in the second case the number of features depends on the number of the anchor words (7 or 26 in our setup).

5.2.3 Results

The results for Continuous Bag of Words (CBOW) word embedding models, produced with 10-fold stratified cross-validation, are presented in Table 5.1. Our evaluation metric is macro-averaged F1 score.

The labels for the different approaches are the same as above. We also use two baselines. The first one is a very simple frequency-based approach. In it, the only feature fed to the classifier is the absolute difference between corpus frequencies of the target word in the current and the previous year. This value is additionally normalized by dividing it by the corpus frequency of the target word in the full Gigaword. This baseline turned out to be very competitive, with only a few techniques outperforming it (see Table 5.1).

The second baseline is the classic Procrustes alignment method. It does not use any domain-specific anchor words, only the cosine distances between wq vectors in the aligned models. Note that we also evaluate incrementally trained embeddings aligned using Orthogonal Procrustes (top right cell), although this is not a widely used setup: as a rule, either Orthogonal Procrustes or incremental training is used to make the models comparable, but not both at the same time. The reason for choosing this setup was to make the presentation of the baseline results more consistent with the anchor words experiments.

Overall, one can see that more words in the anchor sets is beneficial (even though it means lower average word frequency in the set). The fully automatically

Approach	**Macro F1 score**		
Frequency baseline		0.27	
		Embedding model type:	
	Scratch	**Cumulative**	**Incremental**
Procrustes/cosine baseline	0.15	0.24	0.29
		Basic anchor set	
SimDist	0.27	0.17	0.25
SimSub	0.31	0.26	0.26
RankDist	0.28	0.19	0.23
RankSub	0.26	0.22	0.21
		Expanded anchor set	
SimDist	0.25	0.18	0.23
SimSub	0.35	0.31	0.29
RankDist	0.24	0.20	0.28
RankSub	**0.36**	0.30	0.32
		Automatic anchor set	
SimDist	0.23	0.21	0.25
SimSub	0.30	0.30	0.32
RankDist	0.24	0.29	0.21
RankSub	0.26	0.26	0.26
		Peace anchor set	
SimDist	0.20	0.19	0.23
SimSub	0.28	0.27	0.27
RankDist	0.21	0.25	0.16
RankSub	0.22	0.21	0.27

Table 5.1: Macro-F1 scores for predicting conflict state changes (ternary classification).

collected anchor set (just the nearest neighbors of the word '*conflict*') performed almost on par with the best manually collected anchor set (the expanded version), suggesting that borrowing seed words from external sources might not be necessary. Using $\vec{\boldsymbol{r_{cur}}} - \vec{\boldsymbol{r_{prev}}}$ ('Sub' method) is almost always better than using $\cos(\vec{\boldsymbol{r_{cur}}}, \vec{\boldsymbol{r_{prev}}})$ ('Dist' method). As for the using of either cosine similarities ('Sim' quantification mode) or ranks ('Rank' quantification mode) as $\vec{\boldsymbol{r}}$ values, there does not seem to be a clear winner. We also tried to concatenate similarities and ranks to produce the feature vector of size 52. However, this did not improve the classifier performance. Using peace anchor words instead of conflict anchor words yielded consistently lower results, barely managing to reach the performance of the frequency baseline (arguably, it can be changed by using other peace anchors compiled in a different way).

The best results with the conflict anchor words are shown by the 'scratch' models, always outperforming the Procrustes baseline and mostly outperforming the frequency baseline. It means that when using a local semantic change detection method (looking at particular word neighbors) and for this particular task, it is not beneficial to employ schemes of incrementally updating the models with new data[14] or concatenating new corpora with the previous ones. Our guess for the reason of it is that the models trained from scratch on yearly corpora are more 'focused' on the events happening in this particular year, and thus provide more useful vector representations. Another reason for this behavior may be related to the issues of incremental training discussed in Shoemark et al. (2019) and Schlechtweg, Hätty, et al. (2019): it is difficult to tune optimally the parameters of vocabulary expansion and the amount of new data fed to the model (expressed in the number of epochs).

Note, however, that for the Procrustes alignment baseline, the scratch models were the worst choice for alignment, arguably because they are more different from each other than cumulative or incremental ones (since each model is initialized independently and with independent collection of training texts). In fact, the Procrustes alignment with cosine distance approach worked *best* on incremental models in this task: this is the only variation of Procrustes alignment which managed to slightly outperform the frequency baseline. This is an interesting finding in itself: sometimes using several different methods to make embeddings comparable might be beneficial.

The best macro-F1 score of 0.36 was produced by using scratch-trained embeddings, the expanded conflict word list and the 'RankSub' method which employs subtraction of time-specific anchor rank vectors. In this mode, our domain-specific anchor words approach significantly outperforms both baselines in all types of models. Hamilton, Leskovec, et al. (2016b) report almost perfect accuracy for the Procrustes transformation when detecting the direction of semantic change (for example, the meaning of the word '*gay*' moving away from 'HAPPY' and towards 'HOMOSEXUAL'). However, our task and the nature of data

[14] Note though that the best result of the automatic anchor set is achieved using the incrementally trained embeddings. It is not as good as the expanded anchor set with the scratch models but still hinting that the choice of the method is anchor dependent.

Class	Precision	Recall	F1
Changing-to-peace	**0.13** (0.06)	**0.29** (0.06)	**0.18** (0.06)
Stable-state	0.80 (0.79)	0.58 (**0.82**)	0.67 (**0.80**)
Changing-to-war	**0.17** (0.12)	**0.33** (0.08)	**0.22** (0.10)

Table 5.2: Detailed performance of the best armed conflict state change detection method (results of weighted random guess in parentheses).

are different: the time periods are much more granular (years instead of decades) and we attempt to detect subtle changes in words' associations or connotations (often pendulum-like) rather than full-scale lexicographic 'semantic shifts proper'. Note also that it is certainly possible to apply the anchor words technique to Procrustes-aligned embeddings (this may improve the results even more). But we were interested in discovering whether local change detection methods can yield meaningful results on their own, without any specific alignment (or other way of making embeddings comparable). The answer is clearly positive: analysis of local semantic neighborhood even on non-aligned vector spaces is able to achieve higher scores than calculating cosine similarities between aligned vectors.

Table 5.2 provides the detailed per-class performance of the best method. In parenthesis, we give the performance values for the stratified random guess baseline. If the random guess results are better than our system score, they are given in bold; if our system results are better, they are made bold instead. In absolute F1 score values, detecting stability breaks seems to be more difficult than detecting the stable state. The performance for the '*changing-to-war*' and '*changing-to-peace*' minority classes is far from ideal. However, all of the scores for these classes are significantly better than chance, while the scores for the majority class are on par or even below the random baseline scores.

5.2.3.1 Results on peaceful countries

The experiments above implied training and evaluating only on the location names present in the Armed Conflict Evaluation Test Set. This means that each of these locations was involved in an armed conflict at least once in the time span from 1994 to 2010. But what about countries not involved in armed conflicts? We decided to conduct an additional sanity check experiment to make sure that our approach does not predict armed conflicts for locations which have never experienced any. For that, we compiled an additional test set of 'peaceful locations'. It consists of all country names[15] which:

1. never appeared in the Armed Conflict Evaluation Test Set;

[15] Taken from https://en.wikipedia.org/wiki/List_of_sovereign_states

2. occurred at least 100 times in the Gigaword corpus.

This resulted in 111 peaceful locations with the average per-year corpus frequency of 8 863. This is a bit lower than the average location frequency of the 'conflict locations' which is 17 749; the reason is obviously a large number of states which are rarely mentioned in the news texts. For this sanity check, we applied our best approach (scratch-trained embeddings, the expanded anchor word list and 'RankSub' method) to the yearly diachronic embeddings of these peaceful locations from 1994 to 2010. Since there were no armed conflicts in these countries, the gold label for each instance here was 'stable' (meaning that the country remained peaceful compared to the previous year).

It would make no sense to train a classifier on this data (with all instances belonging to one class), so we trained it on the instances from the Armed Conflict Evaluation Test Set (recall that the majority of them belong to the 'stable' class as well). However, this time the trained classifier was evaluated on the features generated for the peaceful location instances ($111 \times 16 = 1776$ instances total). The classifier predicted the 'stable' class for 60% of these instances, matching the corresponding cross-validated recall value for the instances from Armed Conflict Evaluation Test Set (0.58). This gives us ground to argue that the proposed method is able to tell stable locations from those with changing armed conflict states, independent of whether the location ever experienced an armed conflict at all. It is also general enough to perform equally well on the totally unseen data.

Still, tracing actual real-world events by detecting short-term cultural semantic changes in diachronic word embeddings is a difficult task. Calculating cosine similarities between word representations in Procrustes-aligned time-specific word embedding models works well for large-scale shifts observed over decades or even centuries, like in Hamilton, Leskovec, et al. (2016b). However, as we can see, sometimes it can be outperformed by a simpler local method not requiring any alignment (our proposed method of manually selecting a couple of dozens of domain-specific 'anchor words' and then measuring word dynamics in relation to them).

Even the performance of this method is not entirely satisfactory from a practical point of view, achieving a macro F1 measure of only 0.36 on the task of ternary classification of armed conflict state changes. But recall that in this chapter we do not try to build a state-of-the-art system for this particular task (or for event detection in general). Instead, we study what sorts of referential semantic change related information is captured by diachronic word embeddings, and what are the most robust ways to extract this information.

We also remind the reader that incremental updating of word embeddings was not beneficial in this case. However, in other tasks related to semantic change detection it can yield promising results, as we show in the next section 5.3.

5.3 Tracing shifts in semantic relations

Methods based on distributional semantic models (static word embeddings in particular) can be used not only to trace diachronic semantic change in *single words*, but also the temporal dynamics of semantic *relations* between pairs of words (see our discussion of this topic above in Section 3.4 and in the Introduction). Recall that this problem is somewhat similar (but not identical, see the next subsection 5.3.1) to a number of previous formulations: 'temporal co-references' in Tahmasebi, Gossen, et al. (2012), 'temporal correspondences' in Zhang et al. (2015), and 'temporal word analogies' in Szymanski (2017) and Tahmasebi, Borin, and Jatowt (2018). Since the relations we deal with are semantic in their nature, we again assume that their diachronic change falls within the scope of diachronic semantic change modeling, although admittedly in a non-mainstream form.

Manually annotated data on armed conflicts in the world (described earlier in this chapter) can be used to estimate the ability of the existing methods to deal with this task. As we show in this section, the necessary prerequisites for achieving decent performance here are incremental updating of the embeddings with new training texts and expanding the models' vocabulary in the course of this process.[16]

5.3.1 Formulating the task

It is well known that news texts can be predictive of active violent groups (Greenawald et al., 2018). A violent group and a geographical location in which this group is active are linked with a specific type of semantic relationship (similar in principle to hyponymy, meronymy, etc, with the exception that this relationship is not annotated in WordNet). These relationships arguably are also manifested in distributional semantic models trained on large news corpora. When the relationships change (emerge or vanish), embedding space changes as well. This is what we explore in this section.

Consider the following task: given that in 2003, '*Kashmir Liberation Front*' and '*ULFA*' groups were involved in armed conflicts in India, and '*Lord's Resistance Army*' in Uganda, predict what entities played the same role in 2004 in Iraq, given the corresponding diachronic word embeddings trained on 2003 and 2004 news texts. According to the UCDP data, the correct answer consists of three entities: '*Ansar al-Islam*', '*al-Mahdi Army*' and '*Islamic State*'. The nature of the task is conceptually similar to analogy reasoning (Mikolov, Yih, et al., 2013), but with the added complexity of diachronic change.

One can argue that a very similar task can probably be addressed without any distributional embeddings, using only corpus co-occurrence data. For example, it is possible to find an armed group active in a given country by calculating what armed group name co-occurred most often with the country name in the target corpus (or vice versa, to find the relevant country by the armed group

[16] Parts of this section were previously published as Kutuzov, Velldal, et al. (2017a).

name). However, such a solution is not applicable in the context of this chapter (even without taking into account that our thesis is about studying distributional embeddings, not other NLP algorithms). The reason is that the 'co-occurrence baseline' would require a given closed set of all words belonging to the target class (either armed group names or country names) to choose from. Without such a set, it is logically impossible to solve the task: each word X in the vocabulary has only one value of its co-occurrence frequency with another word Y, and there is a potentially infinite number of possible semantic relations for which one might want to generate predictions.

If the closed set does exist, it of course helps to solve the task of choosing the most relevant item in this set (if one knows for sure that '*Ansar al-Islam*' is the name of an armed group, one can infer that it co-occurs with '*Iraq*' more frequently than the names of other armed groups). But this is a different task from what we formulated above. In our setup, we do not assume the existence of a stable and immutable set of armed groups (or even countries). The task does imply that we have access to some example instances (like '*Kashmir Liberation Front*', '*ULFA*' and '*Lord's Resistance Army*' above), but the system predictions are not restricted with respect to any closed set. This is an important difference: we are interested in diachronic drift of semantic relations, which means that the elements participating in these relations are fluid as well. We can expect these elements to appear and disappear (as armed groups obviously do). An immutable set of possible choices is not applicable here, as well as any co-occurrence approach based on such a set. For this reason, we do not compare against methods like this.

On the other hand, vector semantic representations and linear operations defined on these representations do provide straightforward ways to generalize and produce predictions not belonging to the initial set of example instances, as we will show below. In Subsection 5.3.2 we will imitate the standard synchronic vector analogies setup, in order to show that word embeddings indeed capture semantic directions like 'location to armed group', not only more well-acknowledged relations like 'male to female' or 'past to present'. Further on, in the subsection 5.3.3, we propose to use a linear projection based method to solve these 'analogies', and in 5.3.4 show that it achieves competitive performance. With this in hand, in the subsection 5.3.5 we find that the learned relationship projections are preserved through diachronic word embeddings, and that one can use their dynamics to trace particular armed groups being involved in conflicts. Subsection 5.3.6 evaluates several variants of this approach on the UCDP data.

Note that our focus here is again the event-driven drift in meaning manifested in context variance: it is not the lexicographic senses of the names denoting locations and armed groups that change, but rather their 'perceived image' and typical connotations, as represented in the analyzed texts.

For some researchers, it can be natural to think about temporal analogies in terms of *onomasiological* change (as time goes, another word form comes to express the same concept), unlike many examples in the previous chapters, which mostly focused on *semasiological* processes (as time goes, another concept comes to be expressed by the same word form). However, a closer look shows that in fact this setup fuses both semasiological and onomasiological changes, which seem to be inextricably interlinked here:

- Semasiological aspect: the groups and geographical locations themselves are perceived as independent lexical entities which undergo semantic changes. For example, an armed group can gradually drift away from being associated with violence.

- Onomasiological aspect: the concept slot 'armed group in an active conflict' (belonging to the 'armed conflict' semantic frame) is filled with different words in different time periods, as the armed groups appear or disappear. This slot can also remain empty, if there are no inner conflicts in the location. Note that the semantic relation itself still remains the same, but the fillers of its 'an armed group' and 'a government controlling a particular location' slots are changing.

Unlike the lexical replacement studies described in Tahmasebi, Borin, and Jatowt (2018), we deal with instances which are not exclusive at any given time point. The 'armed group' slot is not unique (it exists for each country name) and it can be filled with any number of named entities (including 0).

Our change detection approach can be defined as 'local' to some extent: the linear projections that we learn are mostly based and evaluated on the nearest neighborhood data. But even taking this into account, the whole task is very different from the standard understanding of techniques for local semantic change detection in that its scope is not single words but pairs of typed entities ('location' and 'armed group' in our case) and semantic relations between them. We study how these semantic relations change over time (or remain stable), and how one can infer this information from diachronic distributional representations.

We continue to employ the Armed Conflicts Evaluation Test Set. Almost always, the first participant of the conflict (the *sideA* field of the UCDP metadata) is the government of the corresponding location, and the second participant (the *sideB* field of the UCDP metadata) is some insurgent armed group we are interested in. In cases when the gold data described the conflict as featuring several groups on the *sideB* (several insurgents fighting against the government in one conflict), we created a separate entry for each group. This resulted in a test set of 673 'Location–Insurgent' pairs.

5.3.2 Armed conflicts as linear analogies in a vector space

We first conducted experiments to assess the hypothesis that dense embeddings do contain semantic relationships of the type '**insurgent participant** of an armed conflict to the **location** of this armed conflict'. To this end, we trained a CBOW word embedding model on the full English Gigaword corpus (Parker et al., 2011) with window size 5, minimal count 100, dimensionality 300, 10 negative samples and five epochs. Note we also experimented with the Continuous Skipgram algorithm, but it yielded either comparable or worse results, at the same time being more computationally demanding. For some reasons, it seems that CBOW is often better than Skipgram for learning linear projections: see the same observation for the projection from Ukrainian to Russian model in Kutuzov, Kopotev, et al. (2016).

To intrinsically evaluate this model, we used the very well established Google Analogies Dataset from Mikolov, Yih, et al. (2013), which contains English pairs like 'country to its capital', 'country to its currency', 'city to its state', family relations, etc (we filtered out the sections containing purely morpho-syntactic relations like 'present form of the verb to the past form of the verb'). These proportional analogies were solved with the Vector Offset method also known as '3CosAdd' (Mikolov, Yih, et al., 2013)[17]. When given the analogy $a : b, c : d$ with an unknown d, and a vector model mapping entities to their embeddings, the Vector Offset suggests that d can be found with the equation 5.2. If there is no entity (word) with the exact $\vec{\boldsymbol{d}}$ vector in the given model (and this is most probably the case), the answer is an entity with the highest cosine similarity to the predicted $\vec{\boldsymbol{d}}$.

$$\vec{\boldsymbol{d}} = \vec{\boldsymbol{b}} - \vec{\boldsymbol{a}} + \vec{\boldsymbol{c}} \tag{5.2}$$

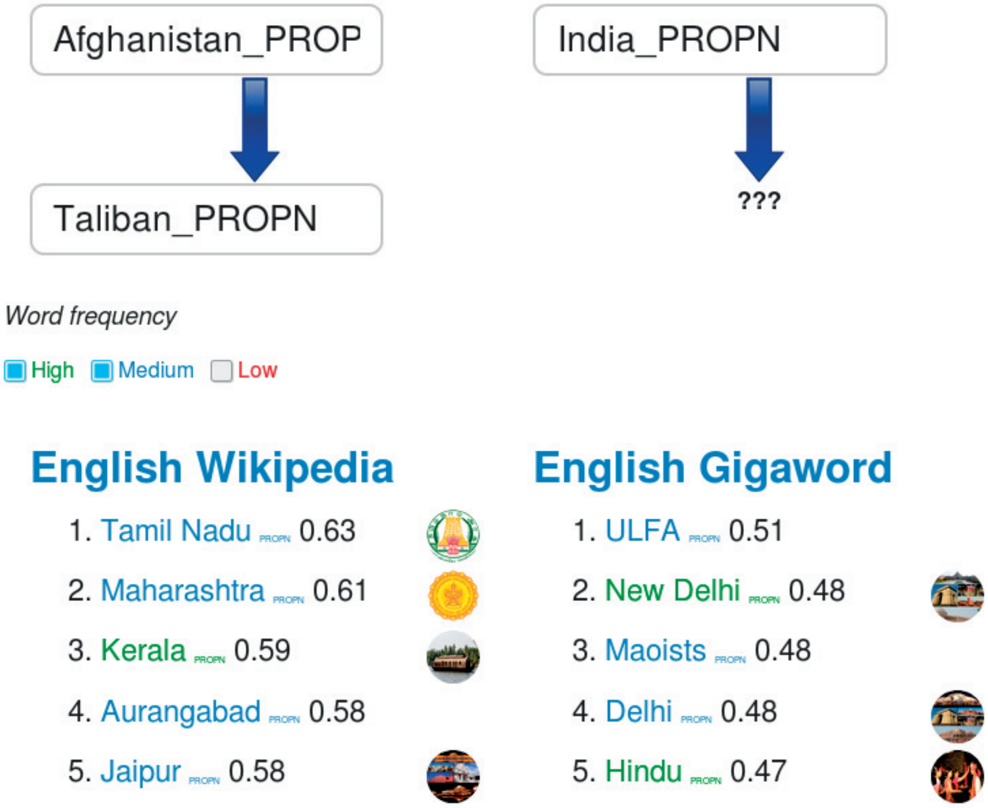
With this test set and the Vector Offset analogy solving method, our CBOW model yielded an accuracy of 70.4%, which is comparable to other English word embeddings (Fares et al., 2017).

The Google Analogies Dataset does not contain 'country to an armed group' pairs. However, hand-picking examples illustrate that word embedding models do capture this information as well. Consider Figure 5.2[18] where the embedding models trained on English Wikipedia and on the Gigaword try to solve the analogy '**Afghanistan** is to **Taliban** is as **India** is to **X**'. **X** is predicted by using the Vector Offset method. Note the differences in the behavior of the two models. The one trained on Wikipedia predicts '*Tamil Nadu*' (one of India states) as an entity being in the same relation to India as Taliban to Afghanistan. Of course Tamil Nadu is related to India, but it is not an armed group, so the answer is incorrect. At the same time, the model trained on news texts (Gigaword) predicts '*ULFA*', the acronym for The United Liberation Front of Assam, which is indeed a militant group banned in India and seeking to establish an independent state of Assam. Thus, the Gigaword model makes it possible to produce the correct prediction in this cherry-picking example, showing that news texts can at least in theory provide enough distributional signal for this kind of tasks.

But what if we evaluate this ability systematically? For this purpose, we created an analogy test set following the same format as the Google Analogies Dataset, but containing the 'Location–Insurgent' pairs from the Armed Conflicts Evaluation Test Set. Typed pairs of countries and armed groups are organized in quadruplets in this dataset version, with 20 942 quadruplets in total. For an example of a quadruplet, see example 1 below, where FARC is an armed group

[17] There also exist other methods to solve vector analogies, see Gladkova et al. (2016) for more details.

[18] Taken from our WebVectors service: http://vectors.nlpl.eu/explore/embeddings/en/calculator/



Afghanistan_PROP → Taliban_PROPN

India_PROPN → ???

Word frequency

☑ High ☑ Medium □ Low

English Wikipedia	English Gigaword
1. Tamil Nadu PROPN 0.63	1. ULFA PROPN 0.51
2. Maharashtra PROPN 0.61	2. New Delhi PROPN 0.48
3. Kerala PROPN 0.59	3. Maoists PROPN 0.48
4. Aurangabad PROPN 0.58	4. Delhi PROPN 0.48
5. Jaipur PROPN 0.58	5. Hindu PROPN 0.47

Figure 5.2: A word embedding model trained on news texts (Gigaword) correctly answering a synchronic armed conflict analogy question (predicting the ULFA armed group active in India).

active in Colombia, while Hamas is an armed group active in Israel. ‘*Hamas*’ is the **X** to be predicted, given three other words in order.

(1) ‘*Colombia FARC / Israel **Hamas***’

The accuracy of our CBOW word embedding model (trained on Gigaword) on this armed conflict test set using the same Vector offset method was much lower than on the Google Analogies Dataset, a mere 3.3%. For consistency with the evaluation below, we also calculated the same accuracy @5 and @10 (that is, whether the correct answer was among top five or top 10 nearest neighbors of the prediction yielded by the model). The resulting accuracy @5 was 11% and the accuracy @10 was 16.2%.

This was expected, since the relations of this kind are much more subtle than those between capitals and countries. Additionally, the names of insurgent groups are mostly low frequency words, compared to the named entities in the Google Analogy test set. This indicates that they have worse embeddings on average, since they were trained on less examples. Finally, an important difference between the datasets is that the Armed Conflicts Evaluation Test Set contains one-to-many relations: some locations map to several armed groups, but the Vector Offset method can output only one answer for each quadruplet (the first nearest neighbor to the calculated vector).

However, we argue these results do not necessarily mean that the relevant relationships are not encoded in the model. They just have to be retrieved by some other means, which we outline in the next subsection.

5.3.3 Learning the armed conflict projection

One intuitive way of improving the performance of solving word analogies (especially with one-to-many data instances) is *learning* from all available related word pairs, instead of solving each analogy separately. In this setup, one trains a supervised model on a train subset of word analogies data and then tests this model on a held-out subset. Drozd et al. (2016) achieve this by simply averaging the offsets between all vector pairs in the training set (they dub this method '3CosAvg', as an allusion to the original '3CosAdd').

We employ the same idea, but cast it as actually learning a *projection matrix* from the embeddings of entities of one type (source) to the embeddings of entities of another type (target). This is done for each section of the dataset, assuming that each of them contains its own type of semantic relations. A similar method has been used for naive translation of words from a L1 language to a L2 language by using monolingual word embeddings for both, and a seed bilingual dictionary (set of one-to-one pairs) (Mikolov, Le, et al., 2013). The theory behind this approach is described in more details in our paper published as Kutuzov, Kopotev, et al. (2016).

Essentially, we train a linear regression which minimizes the error in transforming one set of vectors into another. This amounts to solving d normal equations (where d is the vector size in the embedding model being used, 300 in our case), as shown in equation 5.3:

$$\vec{\boldsymbol{\beta}}_{\boldsymbol{i}} = (\boldsymbol{X}^{\intercal} \cdot \boldsymbol{X} + \lambda \cdot \boldsymbol{L})^{-1} \cdot \boldsymbol{X}^{\intercal} \cdot y_i \tag{5.3}$$

where $\boldsymbol{X}$ is the matrix of source word vectors (input), y_i is the array of the i^{th} components of the corresponding target word vectors (correct predictions), $\boldsymbol{L}$ is the identity matrix of the size d, with 0 at the top left cell, and λ is a real number used to tune the influence of regularization term (if $\lambda = 0$, there is no regularization). $\vec{\boldsymbol{\beta}}_{\boldsymbol{i}} \in \mathbb{R}^d$ is our aim: the vector of dimensionality d such that the dot product of arbitrary source vector and $\vec{\boldsymbol{\beta}}_{\boldsymbol{i}}$ is as close as possible to the i^{th} component of the corresponding target vector. In other words, $\vec{\boldsymbol{\beta}}_{\boldsymbol{i}}$ transforms source vectors into these i^{th} components. After learning $\vec{\boldsymbol{\beta}}$ for each vector component i, we have a linear transformation matrix $\boldsymbol{T} \in \mathbb{R}^{d \times d}$ which is able to transform full vectors into full vectors, thus 'predicting' a target embedding from a source embedding.

This general workflow can in principle be used to solve any word analogies, when one has a set of typed word pairs. In this particular case, our source words

	location→group			group→location		
λ	@1	@5	@10	@1	@5	@10
0.0	0.0	14.6	31.4	8.8	46.7	**70.8**
0.5	0.7	19.0	**35.0**	7.3	49.6	70.1
1.0	**2.2**	**19.7**	32.8	6.6	47.4	66.4
Vector Offset	**3.3**	11.0	16.2			

Table 5.3: Accuracies at different number of nearest neighbors (k) for synchronic projections (time periods not taken into account) from locations to armed groups, and vice versa. The Vector Offset baseline is also reported.

are location names and target words are insurgent armed group names. Below, we apply it to predict insurgents from locations and vice versa in both synchronic and diachronic setups.

To evaluate the accuracy of the resulting projections, we employ leave-one-out cross-validation, i.e., testing the accuracy of predictions on each pair from the existing dataset, after learning the projection matrix on all the pairs except the one used for testing. This means that for the purposes of evaluation, we have to learn the number of projection matrices equal to the number of pairs.

The prediction itself (in the case of predicting an armed group given a location) consists of dot-multiplying the learned transformation matrix $\boldsymbol{T}$ by the location vector from the test pair. It results in the 'armed conflict projection' $\hat{\imath} \in \mathbb{R}^d$. Then, we produce k nearest neighbors of $\hat{\imath}$ in the current distributional model. If the real insurgent group from the test pair was present in these k neighbors, the accuracy for this pair is 1, otherwise it is 0. The overall performance of the system is measured as an average accuracy over all pairs.

5.3.4 Results of synchronic evaluation

While still remaining in the synchronic realm, we first test whether employing the linear projection approach will help us better predict armed groups from their locations. In Table 5.3 we report the average accuracies of linear projections with different values of the regularization hyperparameter λ and considering hits among the 1, 5 and 10 nearest words (these being different values of the k hyperparameter). We had 137 time-independent 'location–insurgent' pairs from the Armed Conflicts Evaluation Test Set in total. For comparison, we again report the performance of the Vector Offset method for this task.

First, we note that relations of this kind are not symmetric: it is much easier to predict the location based on the insurgent (see the right part of Table 5.3) than vice versa (left part of the table). Second, the achieved 'group→location'

Section	# pairs	@1	@5	@10
Capital-Common-Countries	506	0.0	56.5	73.9
Capital-World	4 524	53.0	88.7	93.9
Currency	866	0.0	34.5	41.4
City-in-State	2 467	1.5	22.1	52.9
Family	506	10.5	42.1	52.6

Table 5.4: Accuracies at different number of nearest neighbors (k) for synchronic projections on the pairs from semantic sections of the Google Analogies test set ($\lambda = 1.0$).

results are now very close to the performance of the same linear projection approach with $\lambda = 1.0$ on the Google Analogies test set (after converting it to a set of unique pairs). The results with the Google Analogies are presented in Table 5.4. Note though that predicting locations from armed groups is not our aim in this chapter: we deal with the inverse task of predicting armed groups from location, and this comparison is given only to better illustrate the capabilities of word embeddings to capture relational information.

When considering the 'location→group' results in Table 5.3, one can see that the Vector Offset method is only (marginally) better for accuracy @1, but it is obviously outperformed by the projection method for accuracies @5 and @1 (for any value of λ). It means that employing linear projections leads to better chances for the correct answer to be located in the immediate neighborhood of the system prediction (even if not at the nearest position). Thus, using the proposed workflow does benefit solving word analogies related to armed conflicts.

It also seems that the higher number of pairs in the training set leads (not surprisingly) to a better performance in learning the transformation matrix: see the numbers for the Google Analogies 'Capital-World' section in Table 5.4, which has the largest number of pairs. The 'Capital-Common' section also exhibits high performance, probably because it contains very frequent country and city names which as a rule receive reliable embeddings during training.

Since the amount of armed conflict training pairs is only 137, it comes as no surprise that the results on this dataset (the left part of Table 5.3) are worse than on the Google Analogies (with hundreds and thousands of pairs in its sections). However, with the projection method, the scores are only marginally lower, not substantially worse as is the case with the Vector Offset @1. Upon a closer look, the performance on the conflict data pairs is very similar to the '*Currency*' or '*City-in-State*' sections of the Google Analogies dataset.

We argue that the lower performance on the armed conflict data in comparison to the Google Analogies dataset is additionally explained by the following three factors:

1. the frequency of words denoting armed groups is lower than any of the words in the Google Analogies data set; thus, their embeddings are of lower quality on average;
2. one-to-many relationships in the UCDP dataset (multiple armed groups can act in one location, or one armed group can act in several locations) make learning the transformation matrix more difficult;
3. learning the transformation matrix on the embeddings trained on the whole Gigaword is sub-optimal, since the majority of armed groups were not active throughout all the Gigaword time span.

From the experiment described above, we conclude that semantic relations between locations and insurgents do exist and can be extracted from word embeddings. They are less expressed than the simplistic one-to-one relations like those in the Google Analogies test set, but still can be found using learned projection matrices. The synchronic accuracies are comparable to those on Google Analogies and thus encouraging, especially considering the fact that the armed conflicts in question are not distributed equally across the whole time span of our training corpus. In the next subsection we employ the same approach diachronically.

5.3.5 Learned armed conflicts projections in a diachronic setup

The aim of our diachronic experiments is to find out whether the 'location–insurgent' projections produced from word embeddings trained on one time period will be able to reveal new conflicts that appeared in the next time period (or old conflicts which disappeared). If successful, this would allow us to trace the armed conflict dynamics by calculating whether the armed conflict relation still holds between the name of the country and the name of a group after some time has passed (and the embeddings were updated accordingly). Even more important, this will mean that diachronic embeddings do capture information about changes in semantic relations between lexical entries.

To this end, we first trained incremental Continuous Bag-of-Words models on the yearly subsections of Gigaword texts, starting from the year 1994 to the year 2010 (final Gigaword year). We incrementally updated this same model with new data, saving a separate model after each subsequent year, following Kim et al. (2014). New words were added to the vocabulary of the model if their frequency in the new yearly data conformed to our minimal count threshold of 15 (so called 'vocabulary expansion'), same as in the previous section. Each yearly training session was performed in five epochs, with linearly decreasing learning rate.

At test time, we work with two word embedding models representing two time bins. As a preliminary example, below we evaluate one of such pairs: namely, the model saved after incremental training on the years up to 2000 (M_{2000}), and the model saved after incremental training up to the year 2001 (M_{2001}). Note that we are trying to find insurgents given a location (as was shown earlier, it is generally a more difficult task as compared to 'find a location given an insurgent').

We extract from the Armed Conflicts Evaluation Test Set all the 'location–insurgent' pairs describing the armed conflicts which took place between 1994 and 2000 (this will be 91 pairs total). The projection matrix $\boldsymbol{T}_{2000}$ is learned on the location and insurgent embeddings from M_{2000}. Note that we use all conflicts from 1994 to 2000 to learn the projection matrix, although evaluating only the 2000-2001 pair. The reason behind this is to use all conflict data available from the past in the given point in time (the year 2000 in this case). This is a realistic setup, where one has access to gold annotation for the previous years, but not for the current year (for which the predictions have to be made; in our case, it is the year 2001). In the next subsection 5.3.6, we will thoroughly evaluate both this approach (dubbed 'up-to-now') and its variation where the projection is learned only on the 'salient' conflicts taking place in the most recent year for which we still have gold annotation (the year 2000 in this example). For simplicity, in this subsection we stick to the 'up-to-now' approach.

After the $\boldsymbol{T}_{2000}$ projection is learned, it is applied to the M_{2001} embeddings of the locations which experienced armed conflicts in the year 2001. According to the Armed Conflicts Evaluation Test Set, their number is 47, but after skipping pairs where either the location or the armed group (or both) is missing from the M_{2001} vocabulary, this number lowered to 38. The resulting predicted vectors $\hat{i}$ are evaluated against the real ('gold') armed groups active in the respective locations in 2001, in the same way as in the previous subsection 5.3.4. Ideally performing system is expected to capture all the 'gold' armed groups and thus have the accuracy of 100.

Table 5.5 demonstrates the resulting performance on this particular pair of years taken as an example. These scores reflect how close the predicted vectors were to the actual insurgents. Note that out of 38 armed conflict pairs from 2001, 31 were already present in the previous set of training pairs from 2000 (ongoing conflicts). This explains why the evaluation on all the pairs gives very high results (though still important in confirming that the semantic relations hold after feeding a model with new data).

However, even when evaluated on the seven new conflicts only, the projection performance is encouraging. On the qualitative side, among others, it managed to precisely spot the 2001 insurgency of the members of the Kosovo Liberation Army in Macedonia (accuracy @1), notwithstanding the fact that the initial set of training pairs did not mention Macedonia at all (no armed conflicts took place in this location between 1994 and 2000). Since we did not specifically align the vector spaces, it seems that the incrementally trained embeddings at least partially preserve the existing semantic axis, learned by the model before the new data.

To illustrate the case described above, Figures 5.3, 5.4 and 5.5 plot the 2-dimensional t-SNE (Van der Maaten and Hinton, 2008) projections of location (red) and insurgent (blue) vectors from the 2000, 2001 and 2002 models correspondingly. Black arrows are added to represent the average 'semantic directions' between locations and armed groups. Of course, the original

Conflicts	# of pairs	@1	@5	@10
All	38	44.7	76.3	81.6
Only new	7	14.3	28.6	42.9

Table 5.5: Accuracies of 2000 → 2001 diachronic projection in armed group detection at different number of nearest neighbors (k).

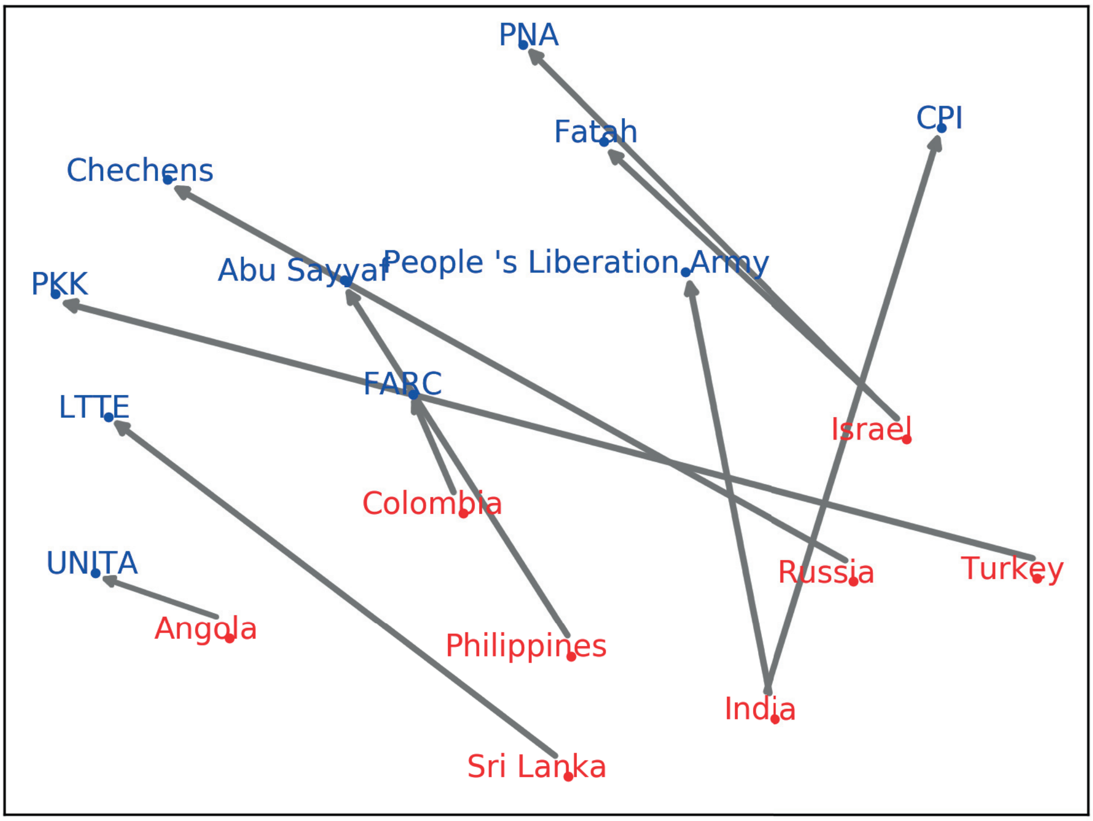


Figure 5.3: Locations (red) and corresponding armed groups (blue) active in 2000; t-SNE projection of high-dimensional word embeddings.

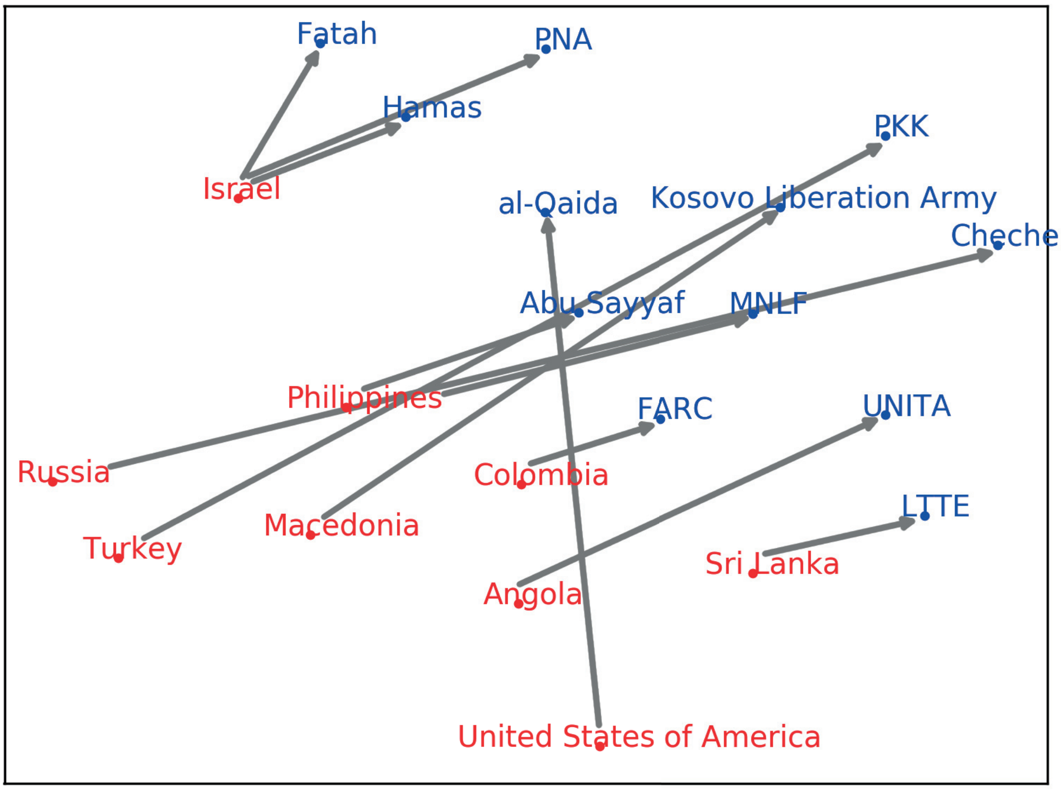


Figure 5.4: Locations (red) and corresponding armed groups (blue) active in 2001; t-SNE projection of high-dimensional word embeddings.

'directions' dwell in the high-dimensional vector space of the original embeddings (300 in our case), but even on the flattened projection these geometrical relations exhibit clear trends within each year. Essentially, most of the arrows are almost parallel, thus ensuring that the averaged 'semantic direction' does make sense. Our projection matrix $\boldsymbol{T}$ is an attempt to learn such a direction from data while minimizing the error.

With the toy evaluation experiment in this subsection, we have shown that a large part of semantic relations within embedding models can survive at least some amount of further incremental training with the new texts that have experienced diachronic change. In the next subsection, we perform a full-scale evaluation of this approach.

5.3.6 Diachronic evaluation of learned projections

In the previous subsection, we tested our approach to predicting future conflicts based on the projection matrix learned from the previous year on the example of one year pair. In this subsection, we systematically evaluate it on the full

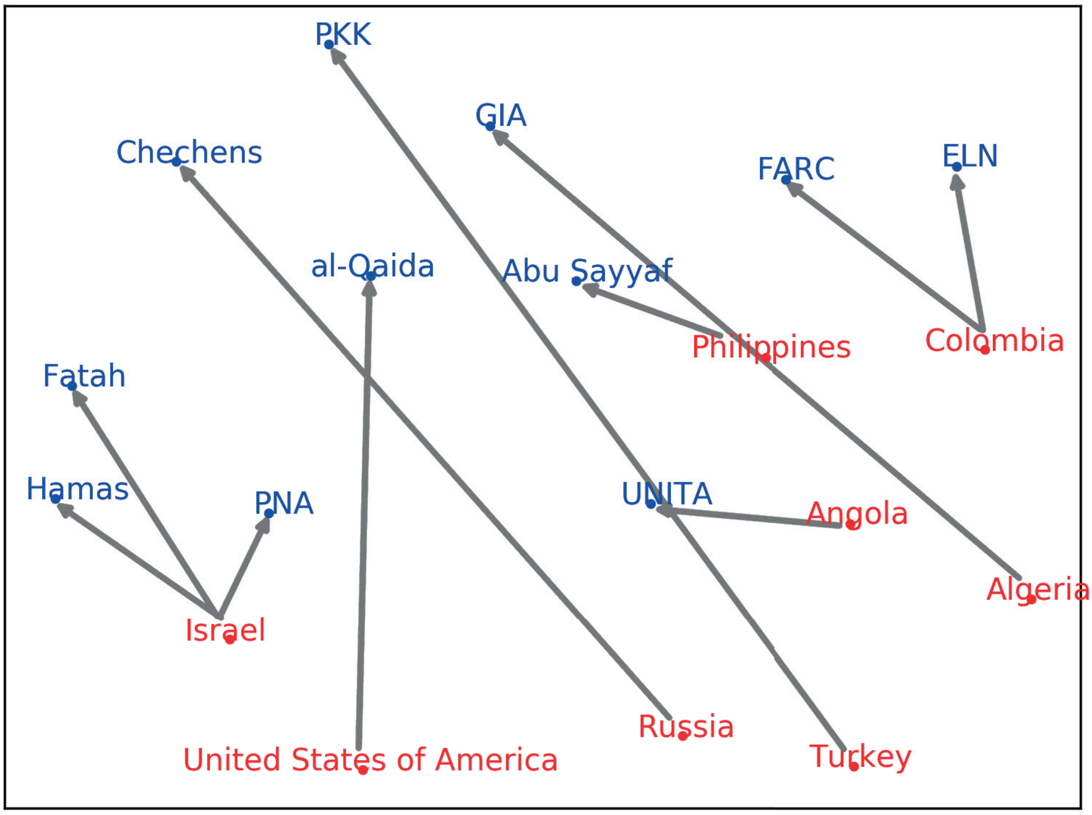


Figure 5.5: Locations (red) and corresponding armed groups (blue) active in 2002; t-SNE projection of high-dimensional word embeddings.

Armed Conflicts Evaluation Test Set for all the years between 1995 and 2010. We did not use the 1994 data, due to its Gigaword sub-corpus being too small and having too many out-of-vocabulary words.

The evaluation metric is the same as before: we calculate the accuracy as the ratio of correctly predicted armed group names for the conflict pairs annotated as active in the Armed Conflicts Evaluation Test Set (gold standard) for this particular year.[19] The final score is the average of these accuracies over all years. Location and armed group names present in the gold standard but missing in the current models (for example, because they were too rare in a particular training corpus and did not manage to reach the frequency threshold) were skipped. In the worst case, 25% of pairs were skipped from the test set; on average, 13% were skipped each year (but see the note below about the 'stable vocab' baseline). At test time, all the entities were lower-cased to eliminate the influence of possible minor differences in spelling of names.

[19] Note that this setup does not ask the used method to predict anything for peaceful locations, which makes the task somewhat easier. We change that in the next Section 5.4.

As before, our word embeddings were incrementally trained on each successive year with vocabulary expansion. We compare them against three baselines:

1. yearly models trained on the concatenation of the texts from the current year and the previous years (hereafter 'cumulative');
2. yearly models trained separately from scratch on the corpora containing news texts from the current year only and then pairwise-aligned using Orthogonal Procrustes (Hamilton, Leskovec, et al., 2016a) (hereafter 'Procrustes');
3. incrementally trained models without vocabulary expansion: they always use the vocabulary from the chronologically first model (hereafter 'stable vocab').

Our initial workflow was to train the linear projections on all the conflict pairs from the past and the current years (hereafter 'up-to-now'). However, this can somewhat decrease the performance, as the information about conflicts having ended several years before might not be strongly expressed in the model after it was incrementally updated with the data from all the subsequent years. For example, the 2005 model hardly contains much knowledge about the conflict relations between Mexico and the Popular Revolutionary Army (EPR) which stopped being active after 1996. Arguably, the direction between these two entities in the 2005 model vector space is very dissimilar from other, more salient 'location–insurgent' pairs.

To test whether this can really influence performance negatively, we additionally conducted an experiment with the projections learned only on the salient pairs (hereafter 'single-year'). In it, we used for training only the pairs active in the last year up to which the model was trained ('current year').

Table 5.6 presents the results for these experiments, as well as for the baselines (averaged across all years). One can see that for the proposed approach (incremental models with vocabulary expansion), the performance of the 'single-year' projections is not worse than that of the 'up-to-now' learning regime. In fact, the former even outperform the latter on the accuracies @1 and @5, while taking less time to learn, because of less training pairs. Our explanation for that is that the single-year projections are more focused on the salient events.

We find that the results of the 'cumulative' baseline are only slightly better than random jitter. This approach performs much worse than the methods which imply making the diachronic embedding comparable: Procrustes and incremental training. This is precisely because the cumulative models are not comparable to each other: they are initialized with different layout of words embeddings in the vector space. This gives rise to formally different directions of semantic relations in each yearly model. The relations themselves are still present in the embeddings, of course, but they are rotated and scaled differently). Note that although using Procrustes-aligned embeddings does yield much better results compared to the cumulative baseline, it still consistently loses out to incremental training. This means that while Procrustes alignment is often

Model type	**up-to-now**			**single-year**		
	@1	@5	@10	@1	@5	@10
			Only in-vocabulary words			
Cumulative	1.7	8.3	13.8	2.9	9.6	15.2
Procrustes	21.2	39.1	47.8	27.4	50.6	58.2
Stable vocabulary	54.9	**82.8**	**90.1**	**60.4**	79.6	84.8
Incremental	32.5	64.5	72.2	42.6	64.8	71.5
			All words, including out-of-vocabulary			
Cumulative	1.5	7.4	12.2	2.5	8.5	13.4
Procrustes	14.5	27.0	32.9	21.6	39.9	46.1
Stable vocabulary	20.8	31.5	34.2	23.0	30.3	32.2
Incremental	28.1	56.1	**62.9**	**37.3**	**56.7**	62.6

Table 5.6: Average accuracies of predicting next-year insurgents based on location vectors and projections learned from the previous year.

reported to outperform incremental training in lexical semantic change detection (Shoemark et al., 2019; Schlechtweg, Hätty, et al., 2019), this is not the case at least when dealing with changes in semantic relations. It seems that incremental training preserves semantic directions in the vector spaces, and does this job better than the attempts to 'restore' these directions by post-hoc alignment.

The results for the 'stable vocabulary' baseline are interesting: this setup uses incrementally updated models without vocabulary expansion (the vocabulary stays the same from the very first model). When tested only on the words present in the test model vocabulary ('Only in-vocabulary words' part of Table 5.6, at the top), this baseline seems to outperform all the other approaches, including those based on incrementally trained embeddings with vocabulary expansion. This stems from the fact that incremental updating with stable vocabulary means we never add new words while updating our models. Thus, they essentially keep the same original vocabulary inferred from the 1994 corpus. The result is that at test time we skip much more out-of-vocabulary (OOV) pairs than with the other approaches (about 62% in average, in comparison to 13% with other model types). Thus, the projections are in fact tested only on a minor part of the test sets (arguably on the easiest, most frequent words).

Of course, simply skipping the larger part of the data under analysis would be a major drawback both for real-life applications and for our task of probing diachronic word embeddings for semantic change information. So the 'stable vocabulary' baseline is not really plausible. For comparison, Table 5.6 additionally provides the accuracies for the setup in which all the pairs are considered ('All words, including out-of-vocabulary' part of the table, at the bottom). In this case, for the pairs with OOV words, the accuracy is set to 0 (not skipped), implying that the system was not able to predict anything. Our approach with vocabulary expansion and other baselines are not much affected by this change: their performance stays almost the same, dropping only marginally (except Procrustes, which suffers somewhat more). But for the 'stable vocabulary' baseline, the scores drop drastically. As a result, after omitting this non-plausible 'stable vocabulary' baseline, incremental training with vocabulary expansion consistently and significantly outperforms all its competitors (including Procrustes alignment) as measured by the average accuracy across all years. We provide the full table of per-year accuracies for this method together with their standard deviations in the Appendix B.

With these experiments, we showed that:

1. Diachronic word embedding-based methods can be used to trace not only semantic shifts in single words, but also changes in typed relations between word pairs.
2. For this particular task, incremental training of embedding models is more useful than alignment with the orthogonal Procrustes transformation.

In the next section, we further refine the evaluation setup of diachronic semantic relations detection, find ways to reduce the number of false armed group predictions, and reproduce our results on newer corpora and test sets.

5.4 Forecasting future armed conflicts as diachronic one-to-X analogies

In this section, we use diachronic word embedding-based methods of semantic change detection to actually predict future armed conflicts and armed groups involved in them ('future' here means 'not seen in the training data and chronologically subsequent').[20] In comparison to the experiments in the previous section 5.3, here we significantly reformulate the analogy task as a 'one-to-X' problem, making it more realistic. We find ways to cope with false positives (insurgent armed groups predicted for locations where no armed conflicts are registered this year). Finally, we use newer and larger corpora of news texts and the most recent version of the UCDP dataset.

5.4.1 Why one-to-X?

The issue of linguistic regularity manifested in relational similarity has been studied for a long time. Due to the long-standing criticism of strictly binary relation structure (see, for example, Turney (2006)), *SemEval-2012* offered the shared task to detect the degree of relational similarity (Jurgens, Mohammad,

[20] Parts of this section were previously published as Kutuzov, Velldal, et al. (2019).

et al., 2012). This meant that multiple correct answers exist ('one-to-many' setup), but they should be ranked differently.

Somewhat similar improvements to the well-known Google Analogies dataset from Mikolov, Yih, et al. (2013) were presented in the BATS analogy test set (Gladkova et al., 2016), also featuring multiple correct answers.[21] Our '*one-to-X*' analogy setup extends this by introducing the possibility of the correct answer being 'None'. In the cases when correct answers exist, they are equally ranked, but their number can be different. Overall, particular instances can be either 'one-to-none', 'one-to-one' or 'one-to-many' relations. Thus, this setup is more difficult than simple 'one-to-many', since there may be zero correct answers (hence 'X' in the name of the approach).

5.4.2 Applying one-to-X to armed conflicts

Once again, we rely on the idea that knowing the gold 'location-insurgent' pairs from a time period t can help us to retrieve the correct pairs bearing the same relation from the next time period $t+1$, using word embeddings trained incrementally on these time periods.

We deal with pairs of consecutive years ('2010–2011', '2011–2012', etc.). Our aim is to predict armed conflicts (or their absence) for a fixed set of locations in the year $t+1$. Having the gold armed conflict data for all years, we can train a predictor on the 1^{st} year, and then evaluate it on the 2^{nd} one (simulating a real-world scenario where new textual data arrive regularly, but gold annotation is available only for older data).

We take the gold 'location-insurgent' pairs from the year t (as a rule, there are several dozens of them) and their vector representations from the corresponding embedding model M_t. Then, these vector pairs are used to train a linear projection $\boldsymbol{T} \in \mathbb{R}^{d \times d}$, where d is the vector size of the embedding model employed. Linguistically, $\boldsymbol{T}$ can be seen as defining a 'prototypical armed conflict relation'; geometrically, it can be thought of as the average 'direction' from locations to their active insurgent groups in the M_t vector space.

The problem of finding the optimal $\boldsymbol{T}$ boils down to a linear regression which minimizes the error in transforming one set of vectors into another, and we do it by solving d deterministic normal equations, described earlier in section 5.3. Since the number of data points is small, the operation is fast. But in the case of large datasets, (almost) the same $\boldsymbol{T}$ can be learned via stochastic gradient descent or any other stochastic optimization process.

After $\boldsymbol{T}$ is at hand, one can find the 'armed conflict projection' vector $\hat{i}$ for any location vector $\vec{\boldsymbol{v}}$ in M_{t+1} by transforming it with the learned matrix: $\hat{i} = \boldsymbol{v} \cdot \vec{\boldsymbol{T}}$. In the simplest case, the word with the highest cosine similarity to $\hat{i}$ in M_{t+1} is assumed to be a candidate for an insurgent armed group active in this location in the time period $t+1$. However, a more involved approach is

[21] See also the detailed criticism of analogical inference with word embeddings in general in Rogers et al. (2017).

needed to handle cases when the number of insurgents (correct answers) can be different from 1 (including 0). This approach is described in the current section.

Note again that for this workflow to yield meaningful results, it is essential for the paired models to be comparable. This is why we train the models incrementally, thus ensuring that they share common structural properties.

5.4.3 Corpora and datasets

In this subsection, we describe the training corpora and the armed conflict datasets we employed.

5.4.3.1 Corpora for embeddings

We train word embeddings on two corpora:

1. The English Gigaword news corpus (Parker et al., 2011), spanning the years 1995–2010 and containing about 300 million words per year, with about 4.8 billion total. This corpus was used in the previous section 5.3 and we include it for comparison purposes.

2. The News on Web (NOW) corpus,[22] spanning the years 2010–2019. The time-annotated texts in NOW are crawled from online magazines and newspapers in 20 English-speaking countries. Since our Armed Conflicts Evaluation Test Set covers conflicts only up to 2017, we use the texts up to that year, yielding on average 730 million words per year, with about 5.9 billion total.

Before training the embedding models, the corpora were lemmatized and PoS-tagged using the UDPipe 2.3 English-LinES tagger (Straka and Straková, 2017) (during the evaluation, PoS tags were stripped and words lower-cased). Chains of consecutive proper names (‘*South_PROPN Sudan_PROPN*’) were merged together with a special character (‘*South::Sudan_PROPN*’). This was important to handle multi-word location and insurgent names (consider ‘*Islamic State*’). Functional words were removed.

5.4.3.2 Conflict relation data

This version of the Armed Conflicts Evaluation Test Set comes from the UCDP/PRIO Armed Conflict Dataset (ver. 18.1) (Pettersson and Eck, 2018). Recall that it is manually annotated with historical information on armed conflicts across the world, starting from 1946, where at least one party is the government of a state, and frequently used in statistical conflict research.

The dataset contains various metadata described earlier in this chapter, but we kept only the years, the names of the locations, and the names of the armed groups. For example, the entry ‘`2016: Afghanistan: ["Taliban", "Islamic State"]`’ (serialized here as JSON) means that in 2016, two armed

[22] https://corpus.byu.edu/now/

	Gigaword	**NOW**
Time span	1995–2010	2010–2017
Unique locations	52	42
Unique armed groups	127	78
Unique conflict pairs	136	102
New pairs share (average)	0.37	0.39
Conflict locations share	0.46	0.56
Insurgents per location	1.65	1.50

Table 5.7: Comparative statistics of our armed conflict test sets.

groups were active in Afghanistan: the Taliban and the Islamic State. Again, entities occurring less than 25 times in the corresponding yearly corpora were filtered out, since it is difficult for distributional models to learn meaningful embeddings for such rare words.

We create one such conflict relation dataset for each news corpus; one corresponding to the time span of NOW and another for Gigaword. Table 5.7 shows various statistics across these test sets, including the important 'new pairs share' parameter, showing what part of the conflict pairs in the years $t + 1$ was not seen in the years t (how much new data to guess).

The new NOW dataset features 102 unique 'location-insurgent' pairs, with 42 unique locations and 78 unique armed groups. On average, each year 56% of these 42 locations were involved in armed conflicts, based on the UCDP data. The remaining locations (different each year) serve as negative examples to test the ability of our approaches to detect cases when no predictions have to be made (since there is no armed conflict in this particular time and location, and thus no semantic relation of this type exists for the current location). For the areas involved in conflicts, the average number of active insurgents per location is about 1.5, with the maximum number being 5.[23]

5.4.4 A replication experiment

In Table 5.8, we report the results of replicating the experiments from the previous section 5.3 on both sets. It follows the same evaluation scheme, where only the presence of the correct armed group name in the k nearest neighbors of the $\hat{i}$ mattered, and only locations with armed conflicts were present in the yearly test sets. In fact, such an approach measures not the *accuracy*, but the *recall* @k, without penalizing the system for yielding incorrect answers along with the correct ones, and never asking questions having no correct answer at all (e.g., peaceful locations).

[23] Congo in the year 2017 featured five active armed groups: '*Kamuina Nsapu*', '*M23*', '*CMC*', '*MNR*', and '*BDK*'.

Dataset	**@1**	**@5**	**@10**
Gigaword	0.356	0.555	0.610
NOW	0.442	0.557	0.578

Table 5.8: Average recall of diachronic analogy inference

The resulting performance is very similar on both sets and on the same level with the results from section 5.3, ensuring that the NOW set conveys the same signal as the Gigaword set. However, in the next subsection we make the task more realistic by extending the evaluation schema to the *one-to-X* scenario described above.

5.4.5 Evaluation setup

In our setup in this section, each yearly test set contains all possible locations, but whether a particular location is associated with any armed group (and thus is plagued by a conflict), can vary from year to year. Conceptually, the task of the system is to predict correct sets of active armed groups for conflict locations (in other words, to correctly predict the nodes in a hypothetical semantic graph which are connected by the 'armed conflict' edges to the location node) and to predict the empty set for peaceful locations. For a test year $t + 1$, an 'armed conflict projection' $\hat{i}$ is produced for each location using its M_{t+1} embedding and the learned transformation $\boldsymbol{T_t}$.

The k nearest neighbors of $\hat{i}$ in M_{t+1} become armed group candidates (k is a hyperparameter). We calculate the number of true positives (correctly predicted armed groups), false positives (incorrectly predicted armed groups), and false negatives (armed groups present in the gold data, but not predicted by the system). These counts are accumulated, and for each year standard precision, recall and F1 score are calculated. These metrics are then averaged across all years in the test set. Using false positives ensures that we penalize the systems for yielding any predictions for locations with no armed conflicts at all. Such cases mean that the system was not able to properly compare relational semantic structures of the M_t and M_{t+1} embedding spaces.

5.4.6 Cosine threshold

It is clear that the system described in the previous subsection (dubbed hereafter 'projection baseline') will always yield k incorrect candidates for peaceful areas.

Inspired partially by the ideas from Orlikowski et al. (2018), we implemented a simple remedy to that, based on the assumption that the correct armed groups vectors will tend to be closer to the $\hat{i}$ point than other nearest neighbors. Thus, the system should pick only the candidates located within a hypersphere of a pre-defined radius r centered around $\hat{i}$. r_t can be different for different time periods t. We infer it from the p training conflict pairs from the previous time period by calculating the average cosine distance between the 'armed conflict projections' $\hat{i}$ and 'gold' armed groups

The procedure is shown in Equation 5.4, where g_p is the embedding of the armed group in the p^{th} pair, and σ is one standard deviation of the cosine distances in p, extending the radius to include more correct predictions.

$$r = \frac{1}{p} \sum_{p=0}^{p} \cos\left(\hat{i}_p, g_p\right) + \sigma \tag{5.4}$$

The hypersphere serves as a cosine threshold. It allows us to keep only the candidates which are not farther from $\hat{i}$ than the armed groups in the previous year tended to be. For example, Figure 5.6 shows a PCA projection of the process of predicting armed groups for Algeria in 2014. With $k = 3$, the system initially yielded three candidates ('*AQIM*', '*Al-Qaida*' and '*Maghreb*'), with only the first being correct, according to the gold data. The red circle is a part of the hypersphere with the radius r_{2013} inferred from the 2013 training data. It filters out the wrong candidates (in black), since the cosine distance from the conflict projection $\hat{i}$ (in blue) to their embeddings is higher than the inferred threshold.

Figure 5.7 shows another example where cosine thresholding improves armed group prediction for Yemen in 2011.

5.4.7 Evaluation of future armed conflicts prediction

For the full-scale experiments on all years of NOW and Gigaword, we chose $k = 2$, to be closer to the average number of armed groups per location in our sets. We evaluate the diachronic performance of our system in the setup when the matrix $\boldsymbol{T}_t$ and the threshold r_t are applied to the year $t + 1$.

First, we test the influence of the cosine thresholding technique without taking peaceful areas into account (that is, excluding the possibility of correct 'None' answers). The resulting scores for both datasets are shown in Table 5.9. As expected, in this setup, using the learned threshold is not beneficial for either dataset: it slightly increases the average precision and slightly decreases the average recall, but the resulting average F1 score remains almost the same. It is not able to make serious contribution to precision (to shift F1 significantly), because in this setup there are no entries for which any answer except 'None' is incorrect.

However, once we move to the realistic one-to-X setup and allow the possibility of 'None' answers (in our domain, such answers are correct for areas without armed conflicts in a particular time period), everything changes, as evidenced by Table 5.10. Taking into account peaceful areas, of course does not change the values of average recall (there are no new armed groups to capture). But as for the average precision, using the cosine-based threshold now makes much larger difference for both Gigaword and NOW datasets (and the corresponding embeddings). The precision differences are statistically significant with *t-test*,

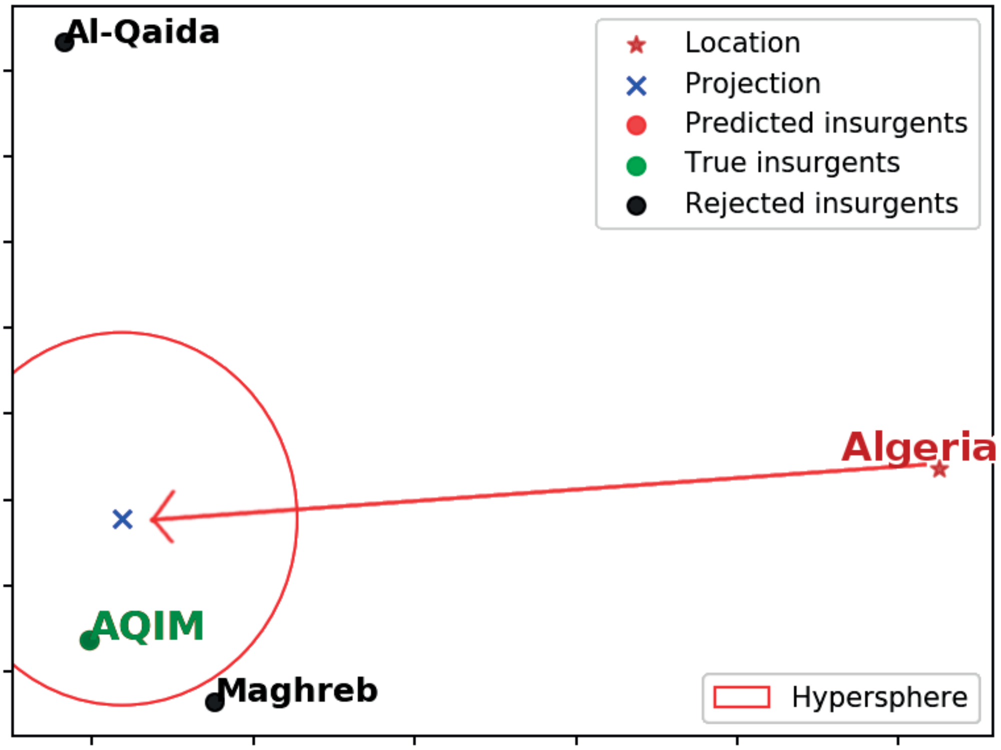


Figure 5.6: Prediction of armed groups active in Algeria in 2014, based on a transformation matrix (red arrow) learned from the 2013 data; 2-dimensional PCA projection.

	Algorithm	**Precision**	**Recall**	**F1**
Gigaword	Projection baseline	0.44	0.51	0.47
	Threshold	0.69	0.41	0.50
NOW	Projection baseline	0.44	0.53	0.48
	Threshold	0.60	0.41	0.48

Table 5.9: Average diachronic performance of armed conflicts prediction with cosine thresholding: testing on conflict areas only.

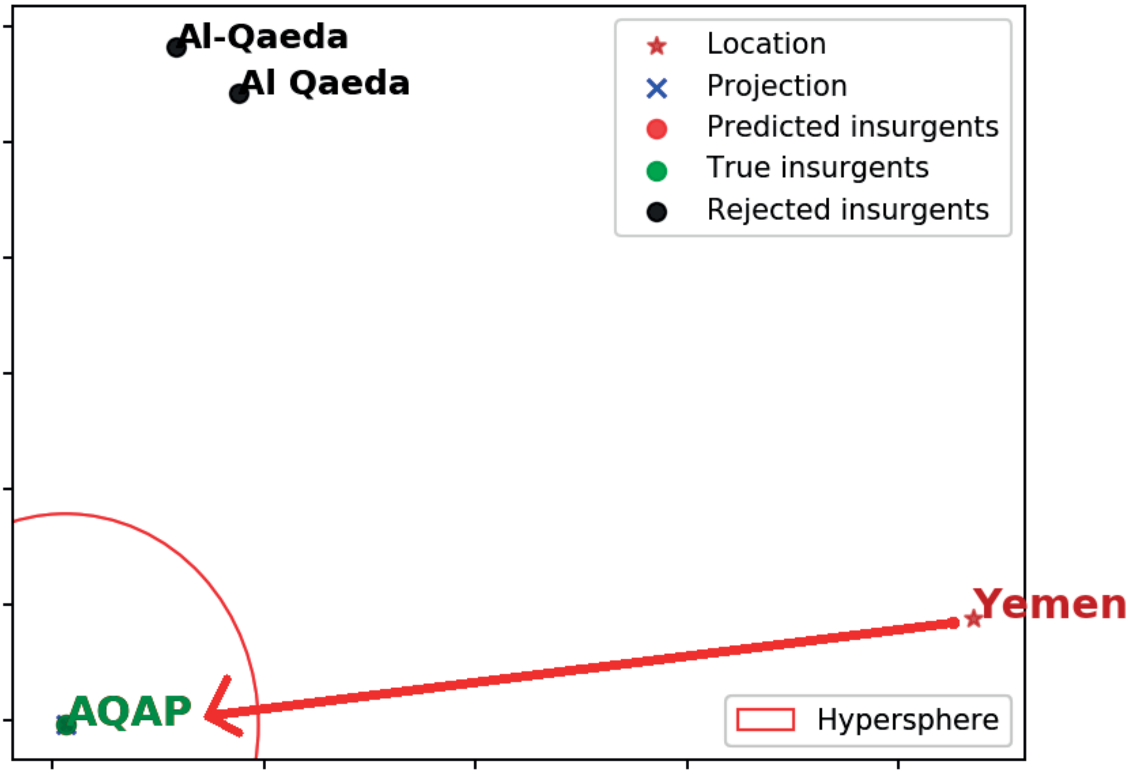


Figure 5.7: Prediction of armed groups active in Yemen in 2011, based on a transformation matrix (red arrow) learned from the 2010 data; 2-dimensional PCA projection.

	Algorithm	**Precision**	**Recall**	**F1**
Gigaword	Projection baseline	0.19	0.51	0.28
	Threshold	0.46	0.41	**0.41**
NOW	Projection baseline	0.26	0.53	0.34
	Threshold	0.42	0.41	**0.41**

Table 5.10: Average diachronic performance of armed conflicts prediction with cosine thresholding: testing on all areas.

$p < 0.05$. Importantly, the integral metrics of F1 consistently improves for the learned cosine threshold ($p < 0.01$), supporting our assumption that this method is effective in cases when 'one -to-zero' entries are real and frequent. The detailed tables with per-year F1 score values for each method and dataset (and their standard deviations) can be found in the Appendix B.

As a sanity check, we also evaluated the same method *synchronically*, that is when $\boldsymbol{T}_t$ and r_t are tested on the locations from the same year t (including

	Algorithm	Precision	Recall	F1
Gigaword	Projection baseline	0.28	0.74	0.41
	Threshold	0.60	0.69	**0.63**
NOW	Projection baseline	0.39	0.88	0.53
	Threshold	0.50	0.77	**0.60**

Table 5.11: Average synchronic performance of armed conflicts prediction with cosine thresholding: testing on all areas.

those lacking any conflicts). In this easier setup, we observed exactly the same trends (see Table 5.11 for the scores).

Thus, our *one-to-X* word analogy task formulation can be applied to the problem of temporal armed conflicts detection based on word embeddings trained on English news texts (from different corpora). A simple thresholding technique based on a function of cosine distance allowed us to significantly improve the relation detection performance, especially for reducing the number of false positives. This approach outperformed the simple projection baseline both with the Gigaword and the NOW news corpora.

The thresholding reduces prediction noise without sacrificing too many correct answers. In our particular case, this helps to more precisely detect events of armed conflicts termination (where no insurgents should be predicted for a location), not only their start. More generally, it means that the system is able to detect cases when the embeddings of a location and an armed group have shifted in such a way in the vector space that it is safe to suppose that they are no longer in the conflict relation. And even more generally, if further supports our hypothesis that diachronic word embeddings can be used to trace subtle temporal changes in semantic relations between words.

We believe that this technique can be employed in a wide variety of applications involving one-to-many or one-to-none relations between linguistic entities. Note, however, that in our experiments, we observed a performance drop if one tries to apply a projection matrix to the embedding model too far away in the future: for example, applying $\boldsymbol{T}_{2000}$ to M_{2010}. This means that over time, incremental updates to the model ‘dilute’ the learned projections, rendering them useless. As a future work, it would be interesting to trace how quickly it happens and analyze the laws governing this deterioration.

5.5 Summary

In this chapter, we introduced the Uppsala Conflict Data Program (UCDP) datasets of armed conflict start and end dates throughout the world. A version of the UCDP Conflict Termination dataset was created and published, convenient for natural language processing tasks (we call it Armed Conflicts Evaluation Test Set). We showed how this data can be used as a source of extra-linguistic indicators useful for probing semantic change detection methods based on diachronic word embeddings, similar to methods of distant supervision. By testing the ability of these methods to detect or predict changes in the real world, we were able to better understand what types of information about changes in lexical semantics is captured by distributional representations. Several different approaches to extract this information were tested and evaluated using the Armed Conflicts Evaluation Test Set. Note that this approach and this particular test set can be used to evaluate semantic change method for any language for which there exists a substantial amount of news texts. Then it is just a matter of translating (or transcribing) the named entities.

News texts are abundant and eagerly cover armed conflicts. This means that beginning or termination of any such conflict is to some extent reflected in the typical contexts surrounding the names of relevant geographical locations and groups in news text published in the corresponding time period. This is the type of information which distributional semantic models efficiently capture.

As an example, one can use such methods in a comparatively simple setup when one measures the temporal drift of a geographical location embedding in relation to conflict domain specific 'anchor words' like '*kill*', '*casualty*', etc. This allows us to detect an armed conflict start or end based only on the analysis of word vector changes which in turn reflect context variance and changes in the referential meaning of a particular named entity. We described such experiments in Section 5.2.

Further on, in Section 5.3, we investigated how incrementally trained diachronic word embeddings can serve as the foundation for systems which are able to trace the dynamics of semantic *relations* over time. This problem is similar to the well-known word analogies task, and is much more difficult and subtle than single-word semantic change modeling, since it involves the analysis of entity tuples (or even triplets or quadruplets).

We considered the task of detecting and predicting armed groups active in particular geographical locations. This is essentially answering the questions like 'Does this semantic relation still hold between the entity X and the entity Y after some time has passed?', where X is, for example, '*India*', and Y is '*United Liberation Front of Assam (ULFA)*'. This setup fuses both *onomasiological* and *semasiological* changes. On the one hand, the stable concept slot 'militant group in an active armed conflict with the national state X' can be filled with different words in different time periods, as the groups appear or disappear (this slot can also remain empty). This is the onomasiological aspect of the task. On the other hand, the groups and geographical locations themselves can be looked at as independent lexical entities undergoing semantic changes (for example, becoming more or less associated with violence). This is the semasiological aspect of the task. Overall, when discussing diachronic changes in *semantic relations*, onomasiological and semasiological shifts seem to be inextricably interlinked.

We addressed this task by learning linear transformations (projections) on diachronic word embeddings. In sections 5.3 and 5.4 it was shown that the projection learning approach significantly outperforms the baselines and can be even applied in the cases of one-to-zero and one-to-many relations. Thus, we proposed a novel model for temporal analogies resolution and redefined the task itself.

The experiments in this chapter involved only one type of relations: that is, armed conflicts. However, the approach of projection learning itself is relation-agnostic. It can be potentially used for any kinds of entities linked by any kind of one-to-X semantic connections which undergo change over time. As shown in this chapter, it can be successfully employed in diachronic tasks as well as in synchronic ones. Provided we possess the relevant corpora, this potentially paves the way to automatically inferring the temporal dynamics of relations between persons and organizations, ideas and technologies, etc.

Analyzing diachronic changes in semantic relations (captured by word embeddings) leads to findings far beyond the usual ‘king is to queen is as man is to woman’ analogy example by Mikolov, Yih, et al. (2013). Such phenomena are more complicated and more interesting, because:

1. the entities can be in one-to-X relations to each other;
2. the entities’ involvement in relations can depend on the time period;
3. the relations themselves can change their form and nature (for example, transforming from one-to-one to one-to-many).

Note that semantic change in this chapter (unlike the previous Chapter 4 and the next Chapter 6) was mostly of referential or ‘world knowledge’ nature. This corresponds to the context variance span on the semantic proximity scale: the words dramatically change their typical contexts without (yet) changing their lexicographic senses. We again argue that such changes are still semantic, although they are different from semantic shifts proper.

For evaluation of diachronic word embedding models, we prepared the Armed Conflicts Evaluation Test Set, converted from the UCDP format to a convenient machine-readable form. The code, test sets and best-performing embeddings from these experiments trained on the English Gigaword and NOW corpora are publicly available (see the last Chapter 7 of the present thesis for the links).

Chapter 6

Contextualized embeddings and semantic change

The previous Chapter 5 employed diachronic word embeddings for practical tasks related to tracing and predicting armed conflicts. In this chapter, we explore contextualized word embedding models based on recurrent neural networks (RNNs) and transformers with regards to their ability to capture lexical semantic change. Here, we test on semantic shifts proper: words acquiring new senses or losing old ones.

In the previous chapters, we used variations of 'static' word embeddings where each occurrence of a word form is assigned the same vector representation independently of its context. Recent contextualized architectures allow us to overcome this limitation by taking sentence context into account when inferring word token representations. The key idea of contextualized embeddings of linguistic entities is that at inference time each word token is assigned a vector representation that is a function of the entire input sentence (Melamud et al., 2016; McCann et al., 2017). It means that these representations are context-dependent: such models will yield different embeddings for one and the same word used in different contexts. Thus, the 'embedding model' is no longer a simple lookup table of word vectors: now even at test time it is a full-fledged deep neural network (trained on a language modeling task), which takes a sequence of words as an input and produces a sequence of context-dependent word vectors as an output. Word vectors themselves are now not fixed: instead, they are learned functions of the internal states of a language model.

However, application of such architectures to diachronic semantic change detection was up to now rather limited, with one paper published in 2019 (R. Hu et al., 2019) and three in the first half of 2020 (Martinc, Montariol, et al., 2020; Martinc, Kralj Novak, et al., 2020; Giulianelli et al., 2020). While all these studies use BERT (Devlin et al., 2019) as their contextualizing architecture, we extend our analysis to ELMo (Peters, Neumann, Iyyer, et al., 2018) and perform a systematic evaluation of various approaches for semantic change detection for both contextualizer architectures. Our experiments show that contextualized embeddings generally outperform previous (for example, static embeddings or Bayesian) approaches, while offering much richer exploration possibilities. We also analyze how ELMo can be used to trace lexical ambiguity changes over time and discuss some critical issues important for applying contextualized architectures in lexical semantic change detection.

Contextualized embeddings allow us to capture word senses in a much more straightforward and efficient way than the context-independent or 'static' word embeddings discussed in the previous chapters. The reason is the higher ‘precision level’: contextualized architectures deal with individual word token representations, not with aggregated word type representations. Since different word senses obviously manifest themselves in different typical token contexts, this is naturally captured by contextualized models, assigning different vector representations to words used in different senses. This is supported by empirical results where contextualized embeddings outperform their static counterparts in word sense disambiguation and word sense induction tasks (Amrami and Goldberg, 2018; Kutuzov and Kuzmenko, 2019).

In the next Section 6.1 we describe ELMo architecture and its relation to word senses in particular. Section 6.2 presents methods for semantic change detection based on contextualized embeddings that we are going to use and evaluate. In Section 6.3 a case study is presented, aimed at finding whether two of these methods capture lexical ambiguity to a level well correlated with human judgment synchronically. For one of the methods, the answer is positive, and we conduct an exploratory experiment to trace lexical ambiguity changes over multiple time bins. However, the core section of this chapter is Section 6.4 where we undertake a systematic evaluation of multiple semantic change detection methods based on contextualized embeddings (both ELMo and BERT). For this, we use the GEMS test set and four test sets from the SemEval-2020 Shared task 1. Finally, Section 6.5 provides qualitative analysis of the results, including some unexpected predictions made by the employed methods. We propose classification and explanation for these issues, as well as possible ways to get rid of them in the future, if need be.

6.1 Embeddings from Language Models (ELMo) as contextualizers

ELMo or ’Embeddings from Language Models’ (Peters, Neumann, Iyyer, et al., 2018) was arguably the first contextualized word embedding model to attract wide attention from the natural language processing community. After advancing the state-of-the-art for a number of NLP tasks, it was awarded Best Paper at the NAACL-2018 conference. Fundamentally, it was based on bidirectional recurrent neural network with two layers.

A surge of other contextualized models has followed, including BERT (Bidirectional Encoder Representations from Transformers) by Devlin et al. (2019), which was based on the Transformer architecture (Vaswani et al., 2017). BERT is essentially a Transformer with self-attention trained on masked language modeling and next sentence prediction. BERT has been shown to outperform RNN-based contextualizers like ELMo in multiple NLP tasks (question answering, natural language entailment, etc), and received the same award at the NAACL-2019 conference. However, ELMo allows faster training and inference than BERT, making it more convenient to experiment with different training corpora and hyperparameters (which is what we do in this chapter). The number of parameters in a typical ELMo model as a rule is only half of that in a typical BERT-base model (57 million versus 110 million), while still offering competitive

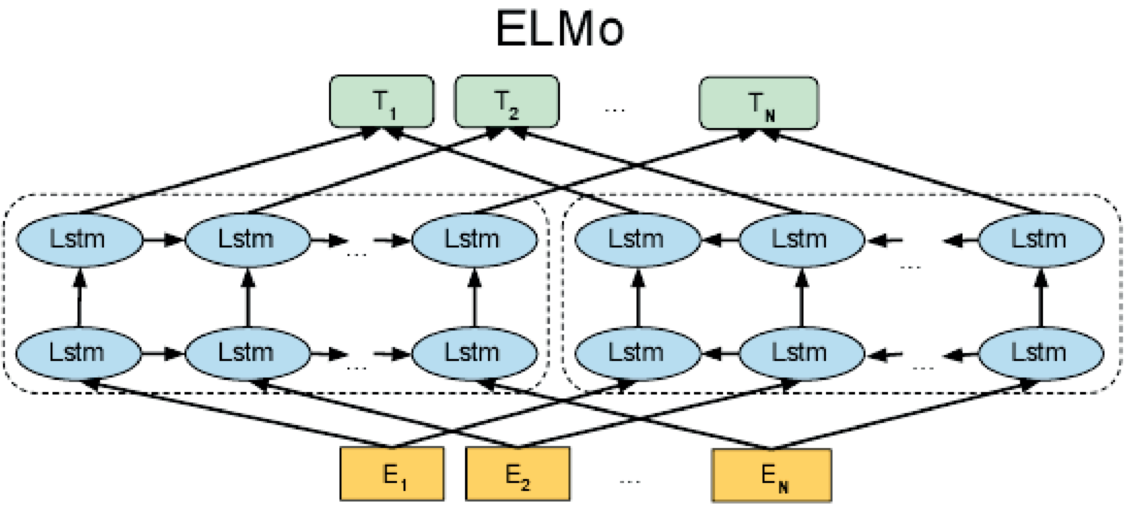


Figure 6.1: ELMo architecture. Image by Karan Purohit, https://medium.com/saarthi-ai/elmo-for-contextual-word-embedding-for-text-classification-24c9693b0045

performance for many tasks. Thus, ELMo remains a very popular algorithm, with pre-trained embeddings available for many languages (Ulčar and Robnik-Šikonja, 2020).

ELMo representations are learned in an unsupervised way through language modeling. The general network architecture consists of a two-layer Bidirectional LSTM (Hochreiter and Schmidhuber, 1997) on top of a convolutional layer which takes character sequences as its input (see Figure 6.1). Peters, Neumann, Zettlemoyer, et al. (2018) have shown that the bottom convolutional layer of ELMo captures word surface forms, the first LSTM layer specializes on syntactic information, while the second (upper) LSTM layer focuses on semantics. For downstream tasks, representations from different layers can be combined in multiple ways, from simple concatenation to learning weighted functions of them (see the next Section 6.2).

ELMo is essentially a pre-trained bi-directional language model. Thus, if one looks at the word representations (embeddings) at the LSTM layers at inference time,[1] one will find that these representations depend on the surrounding words. Of course, representations of one and the same word will still always have something in common, since the convolutional embeddings at the first ELMo layer are deterministic and depend only on the word form (the characters it is composed of). However, as the deterministic embeddings for all words in the input text are passed further to the upper layers, word representations become increasingly context-sensitive. Consider the four English sentences in example 2,

[1] The same is true for the Transformer layers of BERT, but here we focus on ELMo.

all containing an ambiguous word '*bank*' in two different senses:[2]

(2)

1. 'She was enjoying her walk down the quiet country lane towards the river **bank**.' (sense 0)
2. 'She was hating her walk down the quiet country lane towards the river **bank**.' (sense 0)
3. 'The **bank** upon verifying compliance with the terms of the credit and obtaining its customer payment or reimbursement released the goods to the customer.' (sense 1)
4. 'The **bank** obtained its customer payment or reimbursement and released the goods to the customer.' (sense 1)

Obviously, when using a 'static' pre-trained word embedding model, like CBOW (Mikolov, Sutskever, et al., 2013) in the previous chapters, it will return identical vector representations for '*bank*' in all the sentences: simply because the model is essentially a lookup table, mapping words to their embeddings. One can partially override this by representing '*bank*' as an average of the surrounding words' vectors. Actually, averaging representations of context words as a proxy to the sense of one particular word is a long established tradition in word sense disambiguation, starting at least from Schütze (1998). However, this approach is problematic for at least the following reasons:

1. The context words themselves can be ambiguous. Their (also context-dependent) senses are not taken into account.
2. Information about word order in the input data is completely lost, although it can potentially be important for disambiguation.

Contextualized architectures like ELMo and BERT behave completely differently. In example 2, for all four '*bank*' tokens, different representations will be returned, since the context is different in each sentence. However (with good enough embeddings), the '*bank*' vectors in the sentences 1 and 2 will be much closer to each other than to the respective vectors in the sentences 3 and 4 (and vice versa). Thus, in contextualized architectures, the word representations themselves contain information about the particular sense the word was used in in a particular sentence. Also, by looking at these four vector representations, one can possibly infer that the word '*bank*' has two senses, since its token embeddings would likely cluster into two groups. If there are no clearly distinguishable clusters in the set of token representations, this arguably means that the word is mono-semantic (not ambiguous). Thus, using contextualized architectures in theory allows direct access to the information about lexical ambiguity of a particular word.

[2] Examples from the British National Corpus (http://www.natcorp.ox.ac.uk/).

Note that token embeddings belonging to one and the same lexicographical sense can be substantially different in their contexts. This *context variance* (already mentioned earlier in this thesis) is sometimes very systematic and often explained by semantic processes. Thus, strictly speaking, contextualized embeddings do not capture 'pure' lexicographic senses: they rather model what Kilgarriff (1997) called 'senses as clusters of word usages'. We discuss what this issues brings to semantic change detection in Section 6.5.

Another important feature of contextualized architectures is their transferability. One can pre-train a model on a very large corpus (not time-specific), and then use it to produce token embeddings of a target word in other time-specific corpora (they can be much smaller). If the usage contexts for the target word in the corpora are significantly different, the produced embeddings will be different as well, which is important for our topic. Note that this not possible with the traditional static embeddings, where a model (once trained) always produces one and the same representation for a given token. This means that contextualized representations can potentially overcome an important problem in diachronic semantic change detection: historical corpora are often too small in size to train good-quality embeddings solely on them. Additionally, pre-trained embeddings can be fine-tuned on the time-specific corpora. In Section 6.4 below we evaluate both approaches: using pre-trained models 'as is' and after fine-tuning. For ELMo, we additionally test models trained on time-specific corpora only (which would be not feasible for BERT computationally).

6.2 Ways of comparing contextualized embeddings over time

Addressing the problem of polysemy and homonymy was one of the original promises of contextualized embeddings: their primary difference from the previous 'static' generation of word embedding models (Continuous Bag of Words, fastText, GloVe, etc) is that contextualized approaches generate different representations for homographs depending on the context.

Accordingly, we hypothesize that such architectures may provide yet another way to trace and quantify semantic change: by comparing the token representations for a given word in different contexts from different time periods. This whole chapter is dedicated to describing our experiments in this direction. At the time of writing, to the best of our knowledge this is the first attempt at using ELMo for research in diachronic semantic change, and one of the first to use contextualized embeddings at all. In particular, in this section below we outline the contextualized algorithms that we employ for this task.

Comparing contextualized diachronic representations of words over time was studied by R. Hu et al. (2019), by Martinc, Kralj Novak, et al. (2020) and by Giulianelli et al. (2020). Unlike in our case, they used only BERT, not ELMo. Giulianelli et al. (2020) managed to show that contextualized token embeddings do cluster in meaningful groups, corresponding to word senses, and that the analysis of the differences in their distribution over time can help to detect known semantic shifts. However, their empirical results did not outperform previous work in diachronic semantic shift detection: mainly because of using frozen BERT without fine-tuning (see Section 6.4 below for more details). Additionally, their setup poses a conceptual problem of determining the number of clusters (senses) for each target word. This number can be inferred directly from the data, using intrinsic methods like Silhouette score (Rousseeuw, 1987), but it is not very reliable and still requires performing multiple clustering attempts (with different number of clusters) and comparing their scores. Martinc, Kralj Novak, et al. (2020) used the averaging of contextualized token embeddings from BERT, conceptually similar to the PRT measure we describe below. However, their quantitative evaluation was limited to the LiverpoolFC dataset (Del Tredici et al., 2019), which includes only short-term meaning shifts in a particular domain (football). As for R. Hu et al. (2019), they achieved empirical results which outperformed previous work, but only with the help of external data (dictionary sense definitions). See 6.4 below on why we consider their system to be supervised and thus not directly comparable with ours.

How does one use contextualized lexical representations to estimate word meaning change between different (including diachronic) corpora? Below we introduce four possible methods to measure it:

1. Inverted cosine similarity over word prototypes (PRT)
2. Average pairwise cosine distance between token embeddings (APD)
3. Jensen-Shannon divergence between embedding clusters (JSD)
4. Difference between token embedding diversities (DIV)

Given trained contextualized embeddings, a set of corpora and a target word, they produce a score showing how different are the usages of the target word in different corpora (and thus, its meanings, following the distributional hypothesis). Two of the methods (DIV and JSD) additionally produce some estimation of the word's lexical ambiguity along the way. The other two (PRT and APD) are simpler, but, as it turned out, more efficient in practice. All these methods can be used with any contextualized embedding architecture, be it ELMo, BERT or something else. Note though, that using ELMo allowed us to experiment more freely as it has much lower computational requirements than BERT.

We first describe the common part of all four methods. Given two time periods t_1, t_2, two corpora C_1, C_2, and a set of target words, we use a pre-trained neural language model to obtain *contextualized token embeddings* of each occurrence of the target words in C_1 and C_2 and use them to compute a continuous change score. This score indicates the degree of semantic change undergone by a word between t_1 and t_2, and the target words are ranked by its value. This corresponds to the second aspect of modeling of diachronic semantic change that we mentioned in the Introduction: estimating and quantifying the degree of semantic change. This is also how the Sub-task 2 of the SemEval-2020 Task 1 (Schlechtweg, McGillivray, et al., 2020) is formulated.

More precisely, given a target word w and its sentence context $s = (v_1, ..., v_i, ..., v_m)$ with $w = v_i$, we extract the activations of a language model's hidden layers for sentence position i. These embeddings can be collected from the top layer of the used model, averaged over all its layers, or be a product of some weighted function across all layers (we evaluate these options below). If w occurs N times in a given corpus, the N_w contextualized token embeddings collected for w can be represented as the usage matrix $\mathbf{U}_w = (\mathbf{w}_1, \ldots, \mathbf{w}_{N_w})$. The time-specific usage matrices $\mathbf{U}_w^1, \mathbf{U}_w^2$ for time periods t_1 and t_2 are used as input to all the methods of semantic change estimation we describe here.

Contextualized embeddings certainly have their limitations and by no means cover *all* the aspects of lexical ambiguity or semantic change in natural languages. For example, while token embeddings of homonyms (like English '*bank*') will arguably be located far away from each other in the vector space, token embeddings of polysemous words in different but related senses (like English '*paper*' in the senses 'ARTICLE' and 'MATERIAL') will be mutually closer and this might be harmful for attempts to discern these senses. Another source of potential problems is related to frequent cases when a word has one dominant sense and multiple minor senses. As a result, the distribution of word senses in actual usage is very skewed: it is mostly used in the dominant sense, and the corresponding embeddings are close to each other, while the number of tokens used in other senses is so small that it does not influence the overall score. This will pose problems for the methods which implicitly estimate the word's ambiguity (JSD and DIV). We discuss some of these issues in more detail below in Section 6.5.

Even after acknowledging these potential issues, the evaluation results in Section 6.4 still show that the introduced methods can be successfully used to model semantic change, outperforming previous state-of-the-art approaches. We will now describe these methods.

6.2.1 Inverted cosine similarity over word prototypes (PRT)

Given two usage matrices $\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}$, the degree of change of w is calculated as the inverted cosine similarity between the average token embeddings ('word prototypes') of all occurrences of w in the two time periods:

$$\mathrm{PRT}\left(\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}\right) = \frac{1}{d\left(\frac{\sum_{\mathbf{x}_i \in \mathbf{U}_w^{t_1}} \mathbf{x}_i}{N_w^{t_1}}, \frac{\sum_{\mathbf{x}_j \in \mathbf{U}_w^{t_2}} \mathbf{x}_j}{N_w^{t_2}}\right)} \tag{6.1}$$

where $N_w^{t_1}$ and $N_w^{t_2}$ are the number of occurrences of w in time periods t_1 and t_2, and d is a similarity metric, for which we use cosine similarity. This method corresponds to the standard lexical semantic change detection workflow based on static embeddings produced by Procrustes-aligned time-specific distributional models (Hamilton, Leskovec, et al., 2016b), with the only additional step of averaging token embeddings to create a single vector (a prototype). Since we want the method to produce higher scores for the words that changed more, the inverted value of cosine similarity is used as the prediction[3]. The theoretical bounds of the resulting score are $[-\infty, \infty]$, but in practice its values almost always lie in $[1, 2]$.

6.2.2 Average pairwise cosine distance between token embeddings (APD)

Here, the degree of change of w is measured as the average distance between all possible pairs of embeddings from different time periods, following Giulianelli et al. (2020):

$$\mathrm{APD}\left(\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}\right) = \frac{1}{N_w^{t_1} \cdot N_w^{t_2}} \sum_{\mathbf{x}_i \in \mathbf{U}_w^{t_1},\ \mathbf{x}_j \in \mathbf{U}_w^{t_2}} d\left(\mathbf{x}_i, \mathbf{x}_j\right) \tag{6.2}$$

where d is the cosine distance ($1 - c$ where c is cosine similarity). High APD values indicate a higher degree of semantic change. Note that the computational complexity of this measure grows quadratically with the increase in the number of token embeddings, which can become a problem for words with very high frequencies. To cope with this, one can randomly sample a predefined number of token embeddings from both time bins and calculate pairwise distances only between these sampled instances. The bounds of the resulting score are $[0, 2]$.

6.2.3 Jensen-Shannon Divergence between embedding clusters (JSD)

This measure relies on the partitioning of embeddings into clusters of similar word usages. We first follow Giulianelli et al. (2020) and create a single usage matrix with occurrences from two corpora $[\mathbf{U}_w^{t_1}; \mathbf{U}_w^{t_2}]$. We then standardize it by removing the mean and scaling to unit variance, and follow Martinc, Montariol, et al. (2020) to cluster its entries using the Affinity Propagation algorithm (Frey and Dueck, 2007). Affinity Propagation creates clusters by sending messages between pairs of samples until convergence. It is perfect for our task, since it infers the number of clusters for each word directly from the data, without the need to specify it manually. The clusters arguably correspond to the word's senses (their number can be used as a measure of a word's ambiguity). Finally, we define probability distributions $\mathbf{u}_w^{t_1}$ and $\mathbf{u}_w^{t_2}$ based on the normalized counts of word occurrences from each cluster and compute a JSD score (J. Lin, 1991):

$$\mathrm{JSD}(\mathbf{u}_w^{t_1}, \mathbf{u}_w^{t_2}) = \mathrm{H}\left(\frac{1}{2}\left(\mathbf{u}_w^{t_1} + \mathbf{u}_w^{t_2}\right)\right) - \frac{1}{2}\left(\mathrm{H}\left(\mathbf{u}_w^{t_1}\right) - \mathrm{H}\left(\mathbf{u}_w^{t_2}\right)\right) \tag{6.3}$$

Our JSD score measures the amount of change in the proportions of word usage clusters across time periods. The bounds of the resulting score are $[0, 1]$.

[3] We also tried to use cosine distance ($1 - d$) instead of inverted cosine similarity, but the results were marginally worse.

6.2.4 Difference between token embedding diversities (DIV)

Even without actually predicting the exact number of senses, as in JSD (this is a separate difficult NLP task known as 'word sense induction'), it is still possible to estimate the degree of ambiguity for a particular word and then use it to quantify the degree of semantic change. This can be done by calculating the corpus-specific measure we call 'embedding diversity'.

This method is conceptually similar to the notion of 'semantic density' introduced by Sagi et al. (2009) (see Chapter 3); in Section 6.4 we evaluate the 'semantic density' as one of the baselines. DIV estimates the degree of ambiguity for w in $\mathbf{U}_w^{t_1}$ and $\mathbf{U}_w^{t_2}$. Like the PRT method, it first calculates the 'prototype' embeddings p_{t_1} and p_{t_2} by averaging all token representations of w in each usage matrix. The difference is that after this, the mean cosine distances d between w token embeddings and the prototypical embeddings are calculated, thus producing the 'variation coefficients' for both matrices (time periods). The final metrics is the absolute difference between variation coefficients in t_1 and t_2.

$$\mathrm{DIV}\left(\mathbf{U}_w^{t_1}, \mathbf{U}_w^{t_2}\right) = \left| \frac{\sum_{\mathbf{x}_i \in \mathbf{U}_w^{t_1}} d\left(\mathbf{x}_i, p_{t_1}\right)}{N_w^{t_1}} - \frac{\sum_{\mathbf{x}_j \in \mathbf{U}_w^{t_2}} d\left(\mathbf{x}_j, p_{t_2}\right)}{N_w^{t_2}} \right| \tag{6.4}$$

In other words, given a set of contexts (for example, sentences) where a word w occurs, and a pre-trained contextualized embedding model $\boldsymbol{E}$, we:

1. Generate and store a set of representations taken from $\boldsymbol{E}$ for all occurrences of w in a given corpus. The result is a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times s}$ where n is the number of w occurrences (examples) in the corpus, and s is the embedding size (a parameter of $\boldsymbol{E}$).

2. Compute $\vec{\boldsymbol{c_w}} \in \mathbb{R}^s$ vector by averaging across all rows of $\boldsymbol{M}$. $\vec{\boldsymbol{c_w}}$ is, linguistically speaking, a prototypical representation of w, its 'centroid' across different contexts.

3. Calculate the so called 'variation coefficient' by taking the mean cosine distance between each row of $\boldsymbol{M}$ and $\vec{\boldsymbol{c_w}}$.[4] This coefficient itself can be used as the measure of word's ambiguity. The bounds of the resulting score are $[0, 2]$.

The output of these steps for a given corpus is the mean cosine distance between actual token embeddings of the word w from $\boldsymbol{E}$ on a given corpus, and its average, prototypical vector (centroid) $\vec{\boldsymbol{c_w}}$ in the same corpus. Essentially, this is a measure of how varied or diverse the embeddings of w are in the input data, or the measure of w self-similarity. If w is always used in exactly one sense, then arguably its token embeddings will be close to each other, yielding a very low variation coefficient. On the other hand, for highly ambiguous words,

[4] We also tried using average pairwise cosine distances between all rows instead, with the purpose of mitigating the influence of potential outliers (this setup would be even more similar to Sagi et al. (2009)). However, the performance did not significantly differ from the 'centroid' method, while being much slower. Thus, below we use the 'centroid' method in all experiments.

used in many different senses, the 'prototypical' vector will most probably be nonsensical, being equally far away from each sense (and thus from each real token embedding). This will result in high variation coefficient. The more senses the word is used in, the more 'nonsensical' the centroid embedding will be: a randomly sampled word usage is more probable to be distant from the 'centroid' and this probability increases as the number of senses (usage clusters) increases. The DIV method compares variation coefficients of a target word in different corpora (and possibly with different embedding models) by computing their absolute difference to produce a semantic change score.

If a word sense disappears from usage, we expect the variation coefficient to decrease, signaling a change in word meaning. However, the DIV measure does not attempt to actually induce a sense inventory (or even the number of senses) for a particular word. Thus, this change can be caused either by a sense disappearing or by active senses moving closer towards each other.

Before moving on to systematic evaluation of all these methods in Section 6.4[5], we first describe a case study aimed at finding out whether two of them (DIV and JSD) actually capture lexical ambiguity to a degree that is strong enough to correlate well with human-annotated data in a synchronic setup. We also describe a preliminary experiment which involves large-scale tracing of the changes in lexical ambiguity over time. This is what the next Section 6.3 is about.

6.3 Measuring lexical ambiguity with contextualized token embeddings

6.3.1 Synchronic sanity check

Before turning to diachronic data, we evaluate the ability of contextualized embeddings and our methods to produce predictions corresponding to the human-defined degree of semantic ambiguity (at least to some extent). Since we need a notion of ambiguity, we chose the DIV and JSD methods for this experiment: from the former, we use only the diversity coefficient itself; from the latter, we use only the number of clusters produced by Affinity Propagation clustering technique.

The most obvious source of gold sense-related data is the WordNet lexical database (Miller, 1995). It of course does not contain any diachronic data, but it still can be used for evaluation in a synchronic setup.[6] Let us define the degree of semantic ambiguity for a lemma X as the number of WordNet synsets associated with this lemma: for example, the word '*book*' is highly ambiguous, since it is linked to 11 noun and 4 verb synsets, 15 in total.

[5] We also experimented with the change in mean usage relatedness (Schlechtweg, Schulte im Walde, et al., 2018), but never obtained significant correlation with human ratings.

[6] There is ongoing work aiming to build diachronic WordNet-like ontologies, see, for example, Bizzoni et al. (2019). However, at the current time the generated datasets are not mature enough.

To evaluate our methods, we compute their outputs for a reasonable amount of words in some English corpus, and then calculate the Spearman rank correlation between these values and the number of WordNet synsets for the same words. Positive correlation would indicate that the degree of ambiguity can be inferred from contextualized token embeddings, supporting our hypothesis. Surely, from the purely technical point of view, the number of the WordNet senses is easily retrievable for any word mentioned in WordNet or annotated with its WordNet synset. We use this data only to test our hypothesis that the groupings and diversity of contextualized token representations do correlate with the number of word synsets (which in turn approximates the ambiguity of the word).

The evaluation corpus should be reasonably small (so that inferring token embeddings is not too time-consuming), but at the same time contain many ambiguous words. For these reasons (and for this exploratory case study) we used all the sentences from the Senseval-3 English word sense disambiguation dataset (Mihalcea, Chklovski, et al., 2004). It consists of lexical samples for nouns, verbs and adjectives, features about 450 000 word tokens in 3 593 text pieces, and naturally abounds with polysemous words: each text piece contains at least one ambiguous word (the one which has to be disambiguated) and an unknown number of other ambiguous words around it.

Note that we do not use the Senseval-3 annotation: in this experiment, it serves only as a text collection. WordNet, on the other hand, is used as a source of information about the number of senses for the words occurring in this text collection.

See example 3 for an example of the word '*argument*' from Senseval-3 (this dataset defines it as one of its ambiguous target words), used in a mathematical sense:

(3) 'In some situations Postscript can be faster than the escape sequence type of printer control file. It uses post fix notation, where **arguments** come first and operators follow. This is basically the same as Reverse Polish Notation as used on certain calculators, and follows directly from the stack based approach.'

In example 4, we find the same word in the sense of 'REASON GIVEN FOR OR AGAINST A MATTER UNDER DISCUSSION':

(4) 'Environmental organisations, however, put the emphasis elsewhere. There is no incinerator which completely destroys toxic wastes, Madeleine Cobbing, Greenpeace's toxics campaigner, told me. The crux of our **argument** is that we don't need to have toxic wastes and the fact that plants like Rechem exist, providing easy options for companies, is acting as a disincentive for companies to clean up...'

6.3.1.1 Initial experiment

We trained our own ELMo embedding model on the English Wikipedia dump[7] from October 2019. Its size is about 2.8 billion word tokens. The texts were tokenized and lemmatized with the English UDPipe tagger trained on the Universal Dependencies 2.3 treebank (Straka and Straková, 2017), discarding punctuation marks and lower-casing the resulting output. The same preprocessing was applied to the Senseval-3 texts. The ELMo model was trained[8] for three epochs with batch size 192, on two GPUs. To train faster, we decreased the dimensionality of the LSTM layers from the default 4 096 to 2 048.

Then, this model was used to infer contextualized token embeddings for all content words occurring more than two times in Senseval-3 and associated with at least one synset in the WordNet (7 674 in total). We used the *top layer* representations, since the ELMo authors claim that the top layer 'specializes' in semantic-related tasks (Peters, Neumann, Zettlemoyer, et al., 2018). Also, Ethayarajh (2019) found that upper layers of contextualized models yield more context-specific embeddings than the lower levels. This aligns well with our purpose to find embeddings representative of different word senses. In our systematic evaluation in Section 6.4, we, however, also test averaging the embeddings from all layers and from the last four layers for BERT.

For each word, we computed their corresponding ambiguity scores: the variation coefficients (diversities) from the DIV method and the number of token clusters produced by Affinity Propagation from the JSD method. Then, the Spearman rank correlation between the words' ambiguity scores and the corresponding numbers of synsets in WordNet was calculated. For the sake of comparison, we also re-implemented the 'semantic density' algorithm from Sagi et al. (2009) by training an LSI model on the whole Senseval-3 corpus, inferring a matrix of LSI context vectors for each of the target words in the same corpus (where context vector is a normalized sum of vectors for all words within a 15-token symmetric context window to the left and to the right of the current target word occurrence), and then calculating the mean pairwise cosine distance between all vectors in each context matrix. The resulting 'semantic density' value is essentially a measure of how diverse the word contexts are, and so ideally it should correlate positively with the number of WordNet synsets.

The results of the experiments are presented in Table 6.1. The ambiguity scores produced by the JSD and DIV methods across contextualized embeddings show strong and statistically significant (as measured by the two-sided p-value) correlation with the number of synsets, and thus, with the degree of lexical ambiguity. The 'semantic density' values also seem to be well correlated with lexical ambiguity, although to a lesser degree than the ambiguity scores produced by ELMo DIV.

[7] https://dumps.wikimedia.org/

[8] To train and fine-tune ELMo models in this and further experiments, we used the code from https://github.com/ltgoslo/simple_elmo_training, which is essentially the reference ELMo implementation updated to the recent TensorFlow versions.

Model	# of words	Correlation	p-value
ELMo DIV	7 674	**0.4276**	0.000
ELMo JSD	7 674	0.3448	0.000
Semantic density	7 674	0.3468	0.000
Raw frequency	7 674	0.3772	0.000

Table 6.1: The Spearman rank correlations between the number of WordNet synsets and ambiguity scores or word frequency: all words.

However, it has been shown many times (Dubossarsky, Weinshall, et al., 2017; Dubossarsky, Hengchen, et al., 2019; Hamilton, Leskovec, et al., 2016b) that ambiguity and polysemy are involved in complicated relationships with word frequencies (for example, frequent words tend to have more senses and vice versa). Thus, we have to demonstrate that ELMo ambiguity scores are better in approximating the number of synsets than raw word frequencies (as counted on Senseval-3). Indeed, the correlation between word frequencies and the number of synsets is *also* statistically significant and quite strong: see the bottom of Table 6.1 ('Raw frequency' row).[9] ELMo DIV still gives a stronger correlation, but the difference is only five percentage points (and the number of clusters from the JSD is actually *less* correlated with the number of synsets than the raw frequency), and one can ask whether this is worth the effort, if it is equally possible to predict the number of senses by simply looking at the word frequency.

6.3.1.2 Discarding rare words: frequency-controlled experiment

Recall that the frequency distribution of words in human languages obeys the Zipfian power law (Zipf, 1949). It means that among the 7 674 words we analyze, there should be a huge amount of very rare ones. Since rare words tend to be less ambiguous (arguably most of them are associated with only one WordNet synset, if any), they might be the main reason for the strong correlation between frequency and the number of senses: it simply distinguishes between extremely rare words (including *hapax legomena*) which tend to have less synsets and frequent words which tend to have more synsets.

To shed more light on this, we undertake more controlled experiments excluding very rare words. The average frequency of Senseval-3 words is 32, with the standard deviation of 341 (this is a typical power law distribution). We repeated the experiment described above but this time excluding words with the Senseval-3 frequency less than a pre-defined threshold. The threshold values

[9] Interestingly, frequency values calculated on the English Wikipedia (instead of Senseval-3) did not show a significant correlation with the number of synsets. Arguably, the reason is the specificity of the encyclopedic genre.

Model	# of words	Correlation	p-value
ELMo DIV	293	**0.2923**	0.000
ELMo JSD	293	0.0332	0.572
Semantic density	293	0.0746	0.203
Raw frequency	293	0.0632	0.281

Table 6.2: The Spearman rank correlations between the number of WordNet synsets and ambiguity scores or word frequency: only words with frequency higher than 130.

varied from 0 to 130 (where only the 293 top-frequency words are left), with a step of 10. The results are presented in the Figure 6.2.

Most importantly, the more rare words we discard, the worse is the performance of raw frequency, 'semantic density' and the JSD. Eventually, when the words with a Senseval-3 frequency less than 100 are excluded, the frequency-based correlation fails to achieve any statistical significance or strength. To put it simply, after excluding very rare words, it is next to impossible to tell the degree of word ambiguity from its frequency. This supports our hypothesis that the correlation between word frequency and the number of the WordNet synsets is explained by the long tail of rare words.

Another important observation is that the performance of the ELMo JSD method almost perfectly repeats the frequency plot (being marginally worse). It means that clustering ELMo token embeddings with the Affinity Propagation algorithm does not actually yield clusters associated with word senses. Instead, it seems, the number of clusters consistently reflects word frequency. This might be one of the reason why the JSD semantic change detection method based on Affinity Propagation did not show winning results in the empirical evaluation in Section 6.4 below. As for 'semantic density' from Sagi et al. (2009), it performs on par with raw frequency and JSD when rare words are not filtered out, but once this is done, its correlation with WordNet drops even faster.

In contrast, the correlation for ELMo DIV does not suffer much when calculated after excluding infrequent words. It is weaker than when calculated on all words, but overall is quite robust to the changes in the frequency threshold, never falling below 0.25, and always retaining the perfect p-value of 0. Thus, DIV indeed predicts the word's ambiguity, not simply its frequency. Table 6.2 gives the detailed correlation scores for all four methods for the maximum frequency threshold value of 130. To sum up, it seems that contextualized embeddings and our DIV method do collect ambiguity information beyond simple frequency measures and this approach is still efficient when applied to words from a specific frequency tier, unlike its JSD and 'semantic density' counterparts.

If we can approximate lexical ambiguity (that is, the number of senses) of a

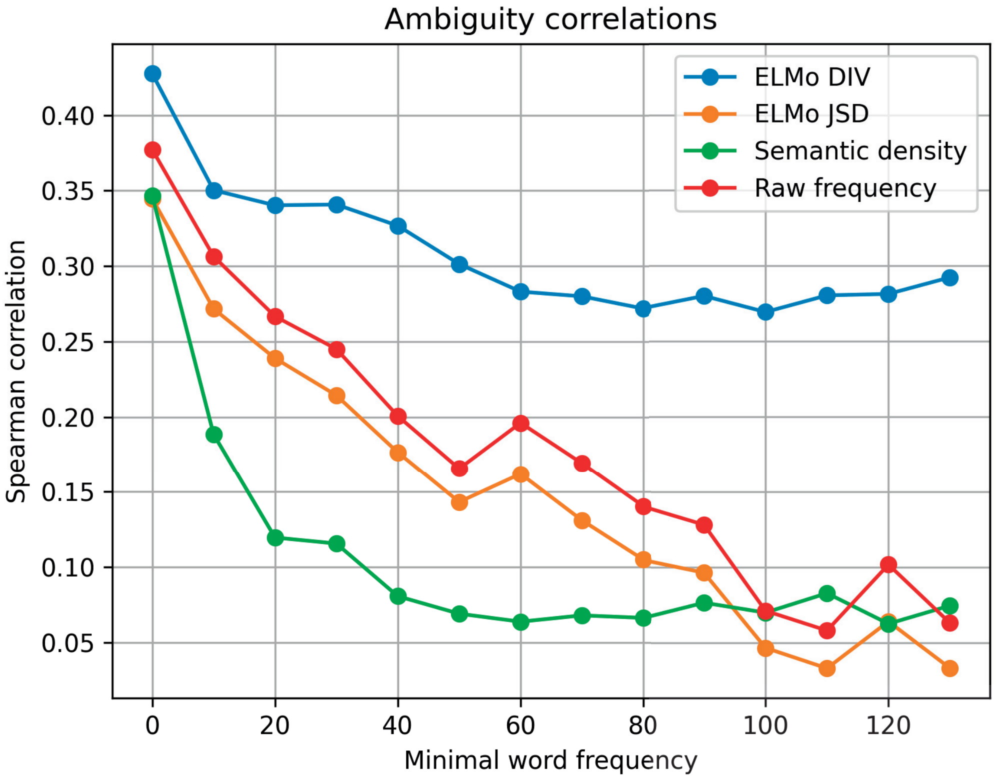


Figure 6.2: Correlation between ambiguity metrics of a word and the number of its WordNet synsets given different frequency thresholds.

word in a given corpus, then large-scale temporal dynamics of ambiguity can also be studied, by simply applying the contextualized ambiguity estimation method to time-specific corpora. We do this in the next subsection 6.3.2, using the DIV method proved to be best for this task in the current section. Note that unlike the rigorous evaluation experiments in Section 6.4, here we do not empirically test the methods themselves: instead, we apply a semantic change modeling method to a large set of words and several time-specific corpora to find out whether we can observe any specific temporal tendencies. In that, this exploratory study is similar to the one we undertook with evaluative adjectives in Chapter 4.

6.3.2 Diachronic ambiguity changes

Following Chapter 4, we use the same diachronic English corpus: namely, the Corpus of Historical American English (COHA). It was pre-processed analogously to the data in the synchronic setup, and split into five time bins corresponding to decades: 1960s, 1970s, 1980s, 1990s and 2000s. We presented the COHA corpus earlier in Chapter 4, and the sizes of each time bin in word tokens were shown in Table 4.1. The full COHA corpus contains texts from the 1810s to the 2000s. We used only the texts created starting from the 1960s, to be consistent with Chapter 4 and with the GEMS dataset used as the ground truth for empirical evaluation in Section 6.4: its time span is from the 1960s to the 1990s.

6.3.2.1 Making ELMo models comparable

As is the case with static embeddings, in order to employ contextualized representations for semantic change detection, these diachronic representations should be first made comparable to each other. With *contextualized* embeddings it is much more difficult to align them using Procrustes transformation and similar approaches. Contextualizing architectures like ELMo and BERT do not define a single weight matrix, unlike the previous generation of architectures that produce static representations. Rather, the full model needs to be applied for each given occurrence of a token, in context, in order to generate a context-dependent representation.

There are currently no standard, established and well-tested methods to align deep neural language models, and it is not immediately clear what can be their possible shared space. Q. Liu et al. (2019) proposed an interesting approach to align contextualized embeddings trained on different languages for the purposes of solving cross-lingual NLP tasks, but this requires a bilingual dictionary. In theory, for diachronic contextualized embeddings trained on different time bins, one can compile a set of monosemous words with extremely stable meaning across the studied time span, but this will immediately raise issues of the principles for such a selection. Schuster et al. (2019) explored methods for cross-lingual alignment of ELMo models in the absence of a dictionary (unsupervised setup), but the performance was worse than in the supervised setup, and the approach is still quite computation-heavy. Overall, inventing and thoroughly evaluating algorithms of contextualized model alignment is an important NLP problem in itself, but out of scope for this thesis. Thus, we do not align our monolingual diachronic contextualized embeddings, but propose and apply two other (much simpler) approaches instead:

1. Contextualized embeddings allow for a conceptually different vectorization setup, where *one and the same single pre-trained model is used* to infer contextualized token embeddings from time-specific corpora. Since the embeddings depend on the context, they are not always identical (as is the case with static embeddings), but yield information about word usage in particular corpora, while at the same time being directly comparable. We will use single pre-trained ELMo embeddings trained either on English Wikipedia or on the full COHA corpus.

2. Another option is to use *incremental training* to make diachronic embeddings comparable. We trained five separate ELMo models incrementally on the COHA time bins: the training of each model except the first one started from the last checkpoint of the previous model (in all cases, we trained

ELMo for five epochs). The softmax layer of all the models used the same vocabulary: the top 100 000 words by frequency across the concatenation of all five diachronic COHA sub-corpora. This setup follows the original idea of incremental training for static word embeddings proposed by Kim et al. (2014): to initialize the model for the time bin t with the weights from the model trained on $t-1$. We already used this approach earlier in the thesis (Section 5.3).

6.3.2.2 Creating samples for frequency tiers

In this exploratory experiment, we aim to find out whether the average degree of lexical ambiguity tends to increase or decrease over time (or stay relatively stable) for words belonging to different frequency tiers. To compare the ambiguity of words in COHA across time, we first found the words occurring in each of the five time bins (45 198 total). This discarded about 500 000 (mostly rare) words for which we would not be able to trace their full evolution from the 1960s to the 2000s anyway. The remaining intersection of the five vocabularies contains both high frequency items (up to 1 million occurrences) and low frequency items (less than 10 occurrences). The distribution of word frequencies is shown in Figure 6.3, with the horizontal axis corresponding to the word's rank in the frequency dictionary (rank 0 is the most frequent word) and the vertical axis corresponding to the word's median frequency across five time bins.

We divided the full intersected vocabulary into three parts, shown on the Figure 6.3 with the red vertical lines. The first part corresponds to the high frequency tier and includes words with a rank up to 10 000; the second part corresponds to the mid frequency tier and includes words with a rank from 10 000 up to 30 000; the third part corresponds to the low frequency tier and includes words with a rank below 30 000. From each part, we uniformly sampled 1 000 random words, thus forming three word lists representing three frequency tiers.[10]

We computed average diversity coefficients (ambiguity scores) of words from each frequency tier in each decade's corpus. The token embeddings were produced with:

1. ELMo model pre-trained on the English Wikipedia,
2. ELMo model pre-trained on the full COHA corpus,
3. the corresponding incremental ELMo model trained on the data up to a given COHA decade.

To avoid the influence of different inference corpora sizes, all the COHA sub-corpora were trimmed to the size of the smallest one: 24 million word tokens (unlike in Chapter 4, we did not discard functional words here). This required removing from 1 to 4 million word tokens per sub-corpus (to do that, we sentence-shuffled each sub-corpus and discarded the necessary number of word

[10] Full word lists are available at https://github.com/akutuzov/elmo_sense/tree/master/data.

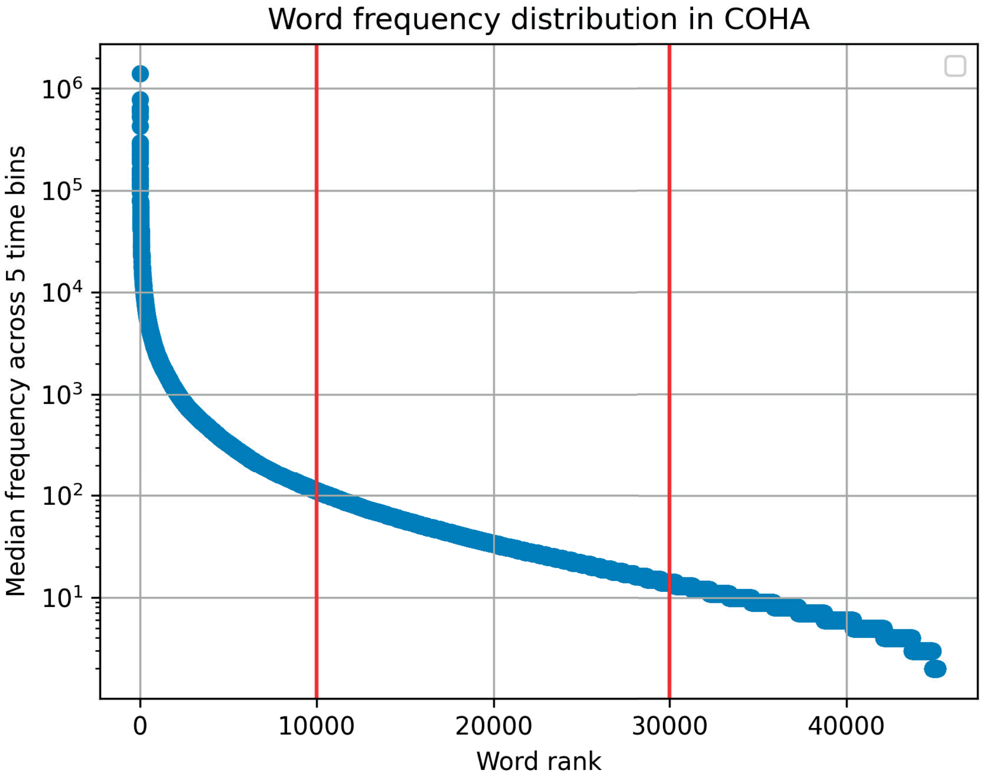


Figure 6.3: Median word frequencies (log scale) for the intersection of COHA sub-corpora vocabularies. Red vertical lines stand for our frequency tier boundaries.

tokens from the end of the sub-corpus). As a result, none of the observations reported below can be caused by fluctuations in per-decade corpus sizes.

6.3.2.3 Results

The results of the DIV calculation are plotted in Figure 6.4 for high-frequency, mid-frequency and low-frequency words. Each line on the plots represents the ambiguity dynamics of a particular word.

No clear tendency with regards to changes in average semantic diversity over time can be inferred from the left and central plots (token embeddings produced by a single model). About three or four words from the high-frequency tier exhibit extremely high and stable diversity values with both models. They are (as expected) mostly very frequent functional words like '*of*' and '*the*' and numerals like '*17*' and '*29*'. They are of course lacking any definite 'sense' and thus are used in very diverse contexts. Another interesting example is the blue line going downwards in the upper part of the mid-frequency plot for the single COHA model (Figure 6.4, central part). This is the word '*thc*'. It was

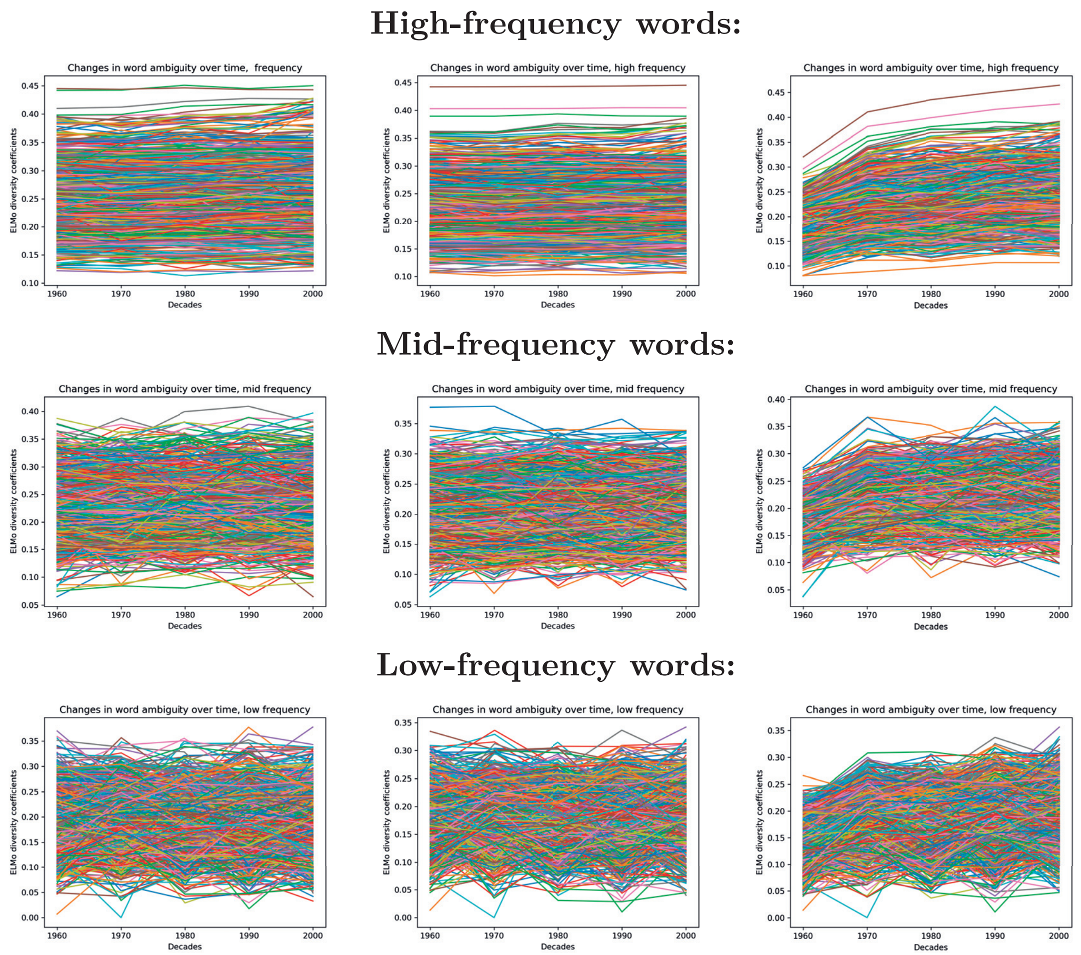


Figure 6.4: Dynamics of ELMo lexical ambiguity (DIV). Left: pre-trained on English Wikipedia; center: pre-trained on the COHA corpus; right: trained incrementally on the COHA corpus.

extremely ambiguous in the 1960s and 1970s, since it usually was a typo for '*the*'. But gradually a specific sense of 'TETRAHYDROCANNABINOL' started to appear, related to marijuana. This led to the decrease of the diversity coefficient for this word, since a cluster of drug-related contexts was formed, where token embeddings were close to each other. This word exhibits a similar evolution with the Wikipedia model, but its diversity values there are not the highest, thus it is visually lost among other words. Overall, the ambiguity dynamics looks like jitter, with no clear trend emerging.

Interestingly, if using token embeddings inferred from *incrementally* trained ELMo embeddings (right parts of the plots), one can clearly observe the tendency for average lexical ambiguity to grow over time. This is most obvious for high-frequency words (including '*of*' and '*the*': the brown and magenta lines at the top of Figure 6.4), but also visible for mid-frequency and low-frequency samples.

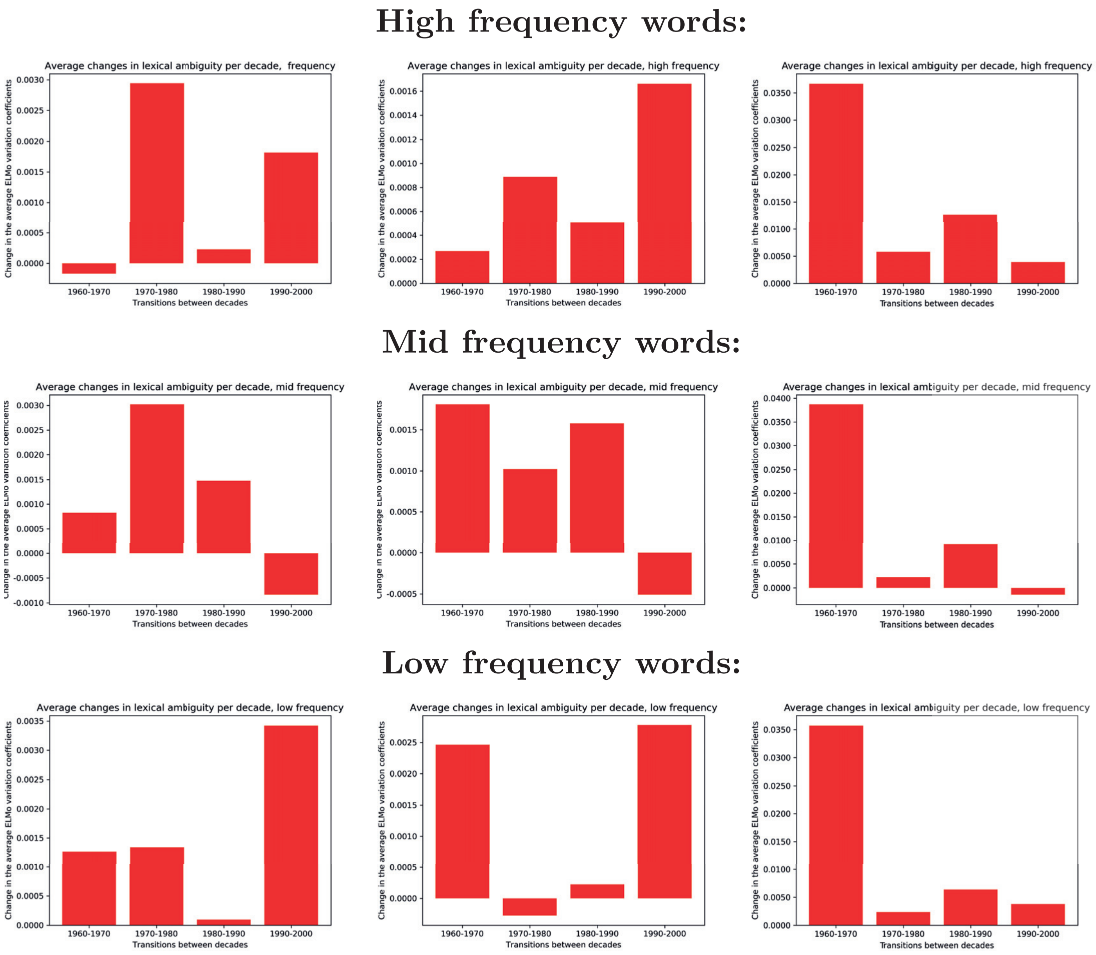


Figure 6.5: Changes in ELMo diversity brought by each decade. Left: Wikipedia model; center: single COHA model; right: corresponding incremental models.

What is the reason for this inconsistency?

It is possible to estimate the change of the degree of ambiguity more formally by simply averaging the pairwise differences between each word variation coefficient in the time bins t and $t-1$ (in other words, their DIV scores as calculated with Equation 6.4). Figure 6.5 describes this for high, mid and low frequency tiers. Each bar there shows the average difference within a pair of decades.

These bar charts and the values behind them show that in fact each new decade does increase lexical ambiguity in COHA. The differences between consecutive diversity scores are almost always positive (or so close to zero that the value is not really significant), independent of the frequency tier or the ELMo model being used. A notable exception is the mid frequency sample in 2000s as compared to 1990s: these words have on average decreased their ambiguity at this time span (as measured by all three models).

However, there is an important difference between the *single* models (Wikipedia and COHA), where token embeddings for all five decades are produced using the same pre-trained weights, and the *incremental* approach where token embeddings for each decade are produced using the model incrementally trained on the texts published up to and including this decade (without 'looking into the future'). The ambiguity differences produced by the single models are very low in absolute values, not exceeding 0.003, while the differences produced by the incremental models are usually twice as high, and for the 1960s–1970s decade pair they are an order of magnitude higher (up to 0.04).

Thus, measuring lexical ambiguity in consecutive time bins with contextualized embeddings incrementally trained on these time bins creates an impression of ambiguity increasing much more than when measured with a single embedding model. It seems that the process of incremental training itself is the reason of diversity increasing: as the model is additionally trained with new data and updates its weights, the representations it produces are becoming more and more diverse.

We ran the same experiment by sequentially employing the same five incremental ELMo models to infer contextualized embeddings for the mid frequency and low frequency words under analysis as occurring in *one* temporal sub-corpus (texts from the 1980s). The contexts and the real ambiguities obviously stay the same in this setup, only the pre-trained embeddings are being changed. If incremental training had no effect at all, the diversity should have stayed approximately the same or fluctuated up and down around some mean value. Instead, in Figure 6.6, we observe that each subsequent model 'sees' more diversity in one and the same collection of texts, with the exception of mid frequency sample for 1990s–2000s, reproducing the dynamics we saw on the previous plots (even including the extremely strong burst of ambiguity in the 1970s compared to the 1960s).

6.3.2.4 Does lexical ambiguity increase over time or not?

Thus, at least part of the 'increase in ambiguity' is due to the confounds of the incremental training process, not to the real linguistic changes (note that we trimmed the decades' corpora to approximately the same size, so this is not a corpus size artifact either). More generally, these phenomena can be seen as an instance of the type of noise effects discussed by Dubossarsky, Weinshall, et al. (2017): observed 'changes' caused not by the trends in the data, but by the peculiarities of the employed algorithms. This systematic noise introduced by continuous training on more data is not unexpected: it was previously described and explained for static embeddings by Schlechtweg, Hätty, et al. (2019) and Shoemark et al. (2019), among others. We now confirmed this for contextualized embeddings as well.

Note also that the DIV scores produced by incremental ELMo embeddings have slightly higher correlation with the changes of *word frequency* from one time bin to another than the DIV scores produced with single COHA or Wikipedia embedding models. The correlation strength is very low in all these cases (even

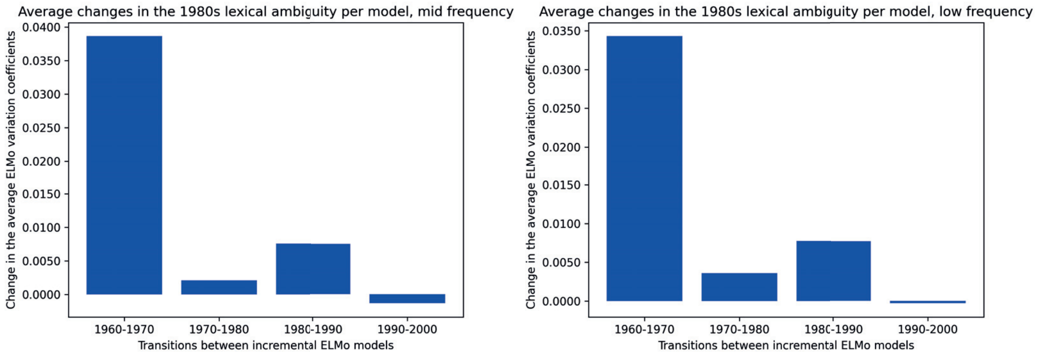


Figure 6.6: Changes in ambiguity for one and the same decade – 1980s – as calculated with five different incremental ELMo models. Left: mid frequency words; right: low frequency words.

in the decade pairs where it is statistically significant, it is only about 0.04 for the single COHA model, about 0.05 for the single Wikipedia model, and about 0.07–0.08 for the incremental models), so the plots above could not be produced from frequencies alone. Still, higher correlation with frequency changes supports the claim we made before: incrementally trained contextualized embeddings are more influenced by the sheer amounts of data (including raw frequency counts), and thus should be used with caution.

Although a large part of the observed increase in variation coefficients is explained by the process of incremental training, we still observe this growing ambiguity even when using single 'frozen' models. In these cases, the effects of incremental training are excluded, but the words' variation coefficients do consistently grow over time, although at a much lower scale than could be induced from looking at the results from the incremental models.

It can be speculated (but only speculated, since it is difficult to draw any grounded conclusions here) that the reason for the ambiguity growing is fast technological and cultural progress, requiring human language (in this case, English) to cover more and more concepts with an inherently limited number of words (neologisms always form only a minor part of the vocabulary in any given period of time). This leads to words from this limited inventory becoming increasingly ambiguous.

We hope to further analyze the results of this exploratory study in future work. For now, it suffices to say that using incrementally trained contextualized embeddings does not look especially promising. This is mostly because they make it difficult to distinguish between 1) representation differences caused by different word usage in two corpora, and 2) representation differences caused by the models being trained on different data. Using one pre-trained model to infer token embeddings for all time bins under analysis avoids this pitfall and allows us to focus on real changes. In the next Section 6.4, we find further support for this hypothesis when evaluating single and incremental approaches on semantic change test sets and show that as a rule incrementally trained contextualized models are outperformed by their single counterparts.

6.4 Empirical evaluation of contextualized methods

In this section, we evaluate the methods described above (PRT, APD, JSD and DIV) on the semantic change test sets from the SemEval-2020 Shared Task 1 (Schlechtweg, McGillivray, et al., 2020).[11]

6.4.1 Description of the task and related datasets

The SemEval-2020 Shared Task 1 challenged its participants to classify a list of target words into stable or changed (Subtask 1) and/or to rank these words by the degree of their semantic change (Subtask 2). The task is multilingual: it includes four lists of target words, respectively for English, German, Latin, and Swedish (several dozen words each). Each word list is accompanied with two historical corpora of varying size, consisting of texts created in two different time periods. Note that two corpora in a pair are not always balanced with regards to their size or the number of occurrences of the target words. This makes the task more realistic. The word lists were manually annotated (in a crowd-sourcing fashion) with respect to the degree of the words' semantic change between the time periods in question. This annotation was held private until the end of the evaluation phase of the shared task.

The shared task organizers additionally provided two baseline methods for both sub-tasks:

1. Normalized frequency difference (**FD**). It first calculates the frequency for each target word in each of the two corpora, normalizes it by the total corpus frequency and then calculates the absolute difference in these values as a measure of change.

2. Count vectors with column intersection and cosine distance (**CNT+CI+CD**). It first learns count-based explicit vector representations for each of the two corpora, then aligns them by intersecting their columns and measures change by cosine distance between the two vectors for a target word.

We evaluate our methods on the Subtask 2[12], with contextualized embeddings based on ELMo and BERT language models. Our evaluation phase submission to the shared task ranked 9th out of 34 participating teams, while in the post-evaluation phase, our submission is the best from those published on the shared

[11] Parts of this section were previously published in Kutuzov and Giulianelli (2020).

[12] We did not specifically focus on the binary Subtask 1; our submission achieved the average accuracy of 0.587 in this track.

task website[13] (but some knowledge of the test sets statistics was needed, see below).

To enrich and diversify our evaluation, we additionally use the GEMS ('GEometrical Models of Natural Language Semantics workshop') test set[14] created by Gulordava and Baroni (2011). It contains 100 English target words. Five human annotators were asked whether each word has changed its meaning from the 1960s to the 1990s (based on the COHA corpus again). It should be noted that GEMS does not explicitly take the number of senses into account in any way: it is just human intuitions about meaning change, independent of whether the number of senses is changing as well. Each word was thus assigned a score on a 4-point scale:

- 0: no change;
- 1: almost no change;
- 2: somewhat changed;
- 3: changed significantly.

We use the average scores as the ground truth. Note that GEMS contains non-lemmatized word tokens (unlike the SemEval-2020 test sets which are lemmatized). In two cases, this leads to two different forms of one and the same word being assigned different scores:

1. '*woman/women*'
   - '*woman*': 2, 1, 0, 0, 0
   - '*women*': 2, 1, 0, 2, 0
2. '*substance/substances*'
   - '*substance*': 2, 2, 0, 1, 1
   - '*substances*': 2, 2, 2, 0, 1

Since we use lemmatized corpora in our experiments, we take the average scores for '*woman/women*' and '*substance/substances*' as the ground truth for '*woman*' and '*substance*', and remove '*women*' and '*substances*' entries. This decreases the total number of words in the test set to 98. For comparison, the SemEval-2020 target word numbers are as follows:

- English: 37 words
- German: 48 words
- Latin: 40 words

[13] https://competitions.codalab.org/competitions/20948

[14] We thank Kristina Gulordava for responding to our e-mail and providing the test set.

	Test set	Median # of w	# of sentences with w	Time span
SemEval	English	208/326	10/8 %	150 years
	German	101/200	3/1 %	118 years
	Latin	427/2 922	22/22 %	2 000 years
	Swedish	254/2 719	1/2 %	89 years
	GEMS	661/923	9/10 %	30 years

Table 6.3: Quantitative characteristics of the lexical semantic change test corpora (w denotes target words). Slashes separate counts for the older and the newer sub-corpus in each pair.

- Swedish: 31 word.

Table 6.3 describes various statistical properties of the SemEval-2020 and GEMS underlying time-specific corpora. As one can see, they vary in the temporal distance between C_1 and C_2 (from 30 years for GEMS to 2000 years for SemEval Latin) and in the median number of test word occurrences in each sub-corpus (SemEval German has the lowest signal here, while SemEval Latin and GEMS have the highest).

In Figure 6.7, we show the absolute frequency and the frequency rank of all target words in each test set (English, German, Latin and Swedish from SemEval-2020, and GEMS). Respective Wikipedia corpora were used for each language. Of course, absolute frequency counts are not directly comparable (for example, Latin Wikipedia is two orders of magnitude smaller than English Wikipedia), but frequency ranks are. English (including GEMS) and Latin test sets contain almost exclusively high frequency entities: the majority of target words have frequency rank higher than 10 000, and only very few are lower than 20 000. The distribution is different for Swedish and (especially) German, where about half of the target words are ranked below 20 000 in the Wikipedia frequency dictionary, about one third is ranked below 40 000, and 11 target words from German are actually extremely rare, with ranks below 200 000. Thus, German and Swedish test sets are more biased towards low frequency words.

For GEMS, our contextualized methods are additionally compared to the static distributional approach originally applied to this test set by Gulordava and Baroni (2011), the SCAN method from Frermann and Lapata (2016) and the frozen BERT APD method which was the best in Giulianelli et al. (2020). Frermann and Lapata (2016) employed a dynamic Bayesian approach, inferring time-specific senses for target words from their contexts, and then calculating their novelty scores. Note that our systems, as well as the one by Giulianelli et al. (2020), do not make any assumptions about the number of senses for the target words. In contrast, the SCAN method uses this number as an additional

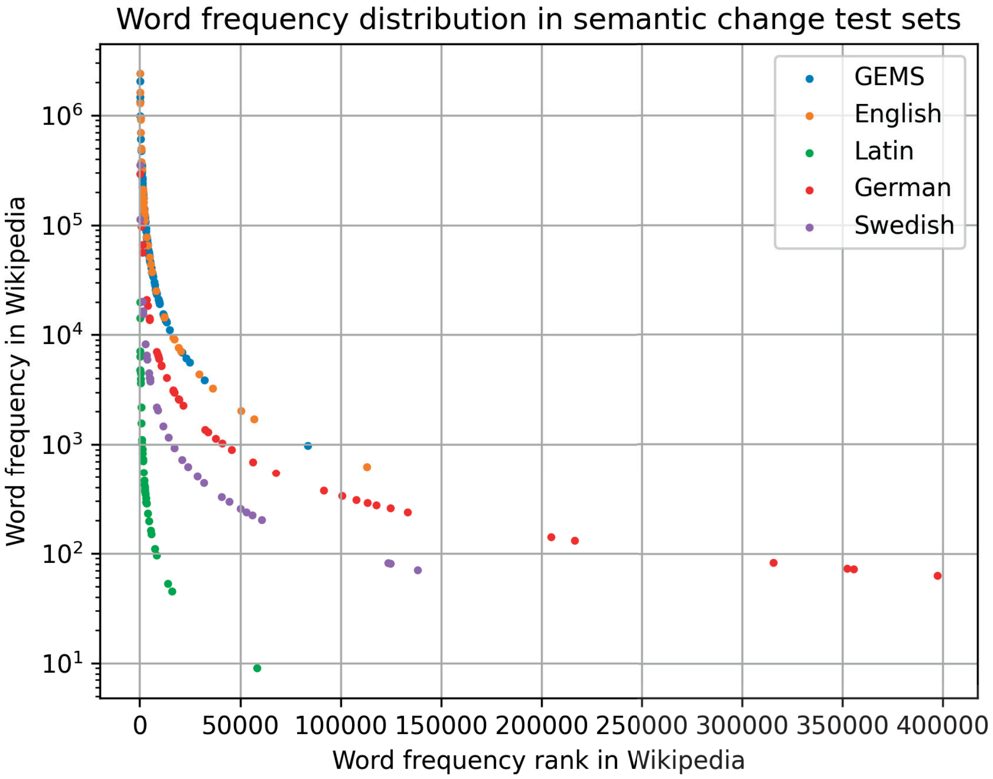


Figure 6.7: Target word frequencies in the semantic change test sets we use (based on the respective Wikipedia corpus for each language). Five German words and one Swedish word omitted from the plot for the sake of visual convenience, since their frequency ranks are below even 400 000.

hyperparameter (it is the same for all words and was set to eight for evaluation on the GEMS test set). The BERT-based system by R. Hu et al. (2019) goes even further and requires *sense embeddings* created by averaging the contextualized representations of target words from the example sentences belonging to different senses of each word. These example sentences (and, consequently, the senses themselves) are extracted from the Oxford dictionary. We argue that this makes it a (semi-)supervised approach, and we do not compare against the GEMS results reported by R. Hu et al. (2019) (Pearson ρ of 0.520 and Spearman ρ of 0.428), since the nature of the data used is completely different, making the systems entirely incomparable.

Martinc, Montariol, et al. (2020) report a Spearman correlation of 0.510 on the GEMS dataset using fine-tuned BERT embeddings with Affinity Propagation clustering and JSD. However, we were unable to reproduce these results, even when using the published code.

6.4.2 Description of our approaches

For each of the 4 languages of the shared task and GEMS, we train 4 ELMo variants:

1. **Pre-trained**, an ELMo model trained on the respective Wikipedia corpus (English, German, Latin or Swedish)[15];

2. **Fine-tuned**, the same as Pre-trained but further fine-tuned on the union of the two test corpora;

3. **Trained on test**, trained only on the union of the two historical test corpora;

4. **Incremental**, two models – the first is trained on the first test corpus, and the second is the same model further trained on the second test corpus.

The ELMo models are trained for three epochs (except SemEval English and Latin **Trained on test** and **Incremental** models, for which we use five epochs, due to the small test corpora sizes), with the LSTM dimensionality of 2 048, batch size 192 and 4 096 negative samples per batch. All the other hyperparameters are left at their default values.

For BERT, we use the *base* version, with 12 layers and 768 hidden dimensions.[16] For English, German and Swedish, we employ language-specific models: *bert-base-uncased*, *bert-base-german-cased*, and *af-ai-center/bert-base-swedish-uncased*. For Latin, we resort to *bert-base-multilingual-cased*, since there is no specific Latin BERT available yet. Given the limited size of the test corpora (in the order of 10^8 word tokens maximum) and BERT's computational requirements, we do not train BERT from scratch and only test the **Pre-trained** and **Fine-tuned** BERT variants. The fine-tuning is done with BERT's standard objective for two epochs (for the English test sets it was trained for five epochs, due to small test corpora sizes). For the English test sets we also tried using the *large* version of BERT with 24 layers and 1 024 hidden dimensions, with only marginal improvements (see Table 6.5).

We configure BERT's WordPiece tokeniser to never split any occurrences of the target words (some target words are split by default into character sequences) and we add unknown target words to BERT's vocabulary. We perform this step both before fine-tuning and before the extraction of contextualized representations.

At inference time, we use all ELMo and BERT variants to produce contextualized representations of all the occurrences of each target word in the test corpora. For the **Incremental** variant, the representations for the

[15] The Wikipedia corpora were lemmatised using UDPipe (Straka and Straková, 2017) prior to training. The punctuation was removed, to better imitate the format of the test corpora in the SemEval-2020 shared task.

[16] We rely on *Hugging Face*'s implementation of BERT (available at https://github.com/huggingface/transformers, version 2.5.0), and follow their model naming conventions: https://huggingface.co/models.

occurrences in each of the two test corpora are produced using the respective model trained on this corpus. The resulting embeddings are of size 12 × 768 and 3 × 512 for BERT and ELMo, respectively. We employ three strategies to reduce their dimensionality to that of a single layer:

1. using only the top layer,
2. averaging all layers,
3. averaging the last four layers (BERT only, since ELMo has only three layers, one of which is purely character-based).[17]

Finally, to predict the strength of semantic change of each target word between the two test corpora, we feed the words' contextualized embeddings into the four methods of semantic change estimation described in Section 6.2. We then compute the Spearman correlation of the estimated change scores with the gold answers. This is the evaluation metric of the SemEval-2020 Task 1's Subtask 2, and we use it throughout our experiments.

6.4.3 Results

In Table 6.4, we report the performance of our contextualized embeddings models on the GEMS dataset, along with the corresponding performance scores taken from Gulordava and Baroni (2011), Fermann and Lapata (2016), and Giulianelli et al. (2020). We also report the performance of the SemEval-2020 baseline methods (Schlechtweg, McGillivray, et al., 2020), the 'semantic density' LSI method from Sagi et al. (2009) and the standard `word2vec` cosine similarity methods. For the latter, we trained CBOW embeddings on the corresponding historical corpora in two different flavors (see Chapter 3):

1. 'Incrementally trained', where the C_2 model was initialized with the C_1 weights (Kim et al., 2014)
2. 'Procrustes-aligned', where the two embeddings were trained independently on C_1 and C_2, and then aligned using the orthogonal Procrustes transformation (Hamilton, Leskovec, et al., 2016b)

We re-implemented the 'semantic density' approach by Sagi et al. (2009) using the Latent Semantic Indexing module from the Gensim library (Řehůřek and Sojka, 2010). Hyper-parameters from the original paper were reconstructed as much as possible, with the notable exception of not using a TF/IDF weighting scheme, since our historical corpora are not separated into documents. Also, in 2020 we are lucky to have much more computational power than back in 2009, and thus we were able to actually calculate density scores on *all* word occurrences, not on a random sample.

[17] We also experimented with the average of layers 5, 6, 7, and 8 but obtained no improvement over the strategies described above. Using summation instead of averaging did not bring improvements either.

Model		**Scores**	
SemEval-2020 baselines			
Frequency-based (FD)		0.068	
Count-based (CNT+CI+CD)		0.256*	
Prior work			
(Sagi et al., 2009)		0.155	
(Gulordava and Baroni, 2011)		0.386*	
(Fremann and Lapata, 2016)		0.377*	
(Giulianelli et al., 2020)		0.285*	
Word2vec cosine similarity			
Incremental models		**0.424***	
Procrustes-aligned		0.235*	
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-train.	0.439*	0.438*	0.406*
+ fine-tuning	0.394*	0.442*	0.417*
ELMo Pre-train.	0.381*	0.365*	–
+ fine-tuning	0.323*	0.332*	–
Trained on test	0.316*	0.293*	–
Incremental	0.414*	0.370*	–
	Pairwise distance (APD)		
BERT Pre-train.	0.203*	0.258*	0.171
+ fine-tuning	0.243*	0.281*	0.214*
ELMo Pre-train.	**0.424***	0.385*	–
+ fine-tuning	0.323*	0.290*	–
Trained on test	0.392*	0.275*	–
Incremental	0.416*	0.388*	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-train.	**0.456***	0.455*	0.405*
+ fine-tuning	0.433*	0.428*	0.431*
ELMo Pre-train.	0.076	0.287*	–
+ fine-tuning	0.225*	0.111	–
Trained on test	0.226*	0.196	–
Incremental	0.035	0.079	–
	Diversity (DIV)		
BERT Pre-train.	0.199	0.258*	0.204*
+ fine-tuning	0.148	0.224*	0.167
ELMo Pre-train.	0.278*	0.267*	–
+ fine-tuning	0.301*	0.314*	–
Trained on test	0.275*	0.138	–
Incremental	0.137	0.213*	–

Table 6.4: The correlations between our predictions and human judgments from the GEMS test set (change between the 1960s and the 1990s).

We report Spearman correlations between our predictions and ground truth scores from the GEMS test set. In all the tables here and below, '*' denotes statistical significance (as measured by the two-sided p-value, $p < 0.05$). Note that the GEMS inter-annotator agreement, measured as an average of pairwise Pearson correlations across five participants, was 0.51 (Gulordava and Baroni, 2011). This value is an approximate upper bound for the systems' performance on this data.

Bold values in Table 6.4 are the best results for *word2vec*, ELMo and BERT. One can see that the baseline and prior work methods are outperformed even by a simple cosine similarity between *word2vec* embeddings incrementally trained on the corresponding COHA corpora. ELMo embeddings pre-trained on English Wikipedia and used with the APD method perform on par with the static embeddings here ($\rho = 0.424$). BERT outperforms both these approaches using the JSD method.

Arguably, some of these approaches (count-based embeddings from Gulordava and Baroni (2011), SCAN, BERT-based and ELMo-based) are not directly comparable, because of different training corpora. For example, the pre-trained ELMo and BERT are different in that our ELMo is trained on Wikipedia only, while the BERT embeddings were pre-trained on a much larger text collection (more than 3 billion words in size, see Devlin et al. (2019) for details). Fremann and Lapata (2016) used the DiAchronic TExt Corpus (DATE), which consists mostly of COHA augmented with 5 million word tokens from other sources, both for training and for inference. Finally, Gulordava and Baroni (2011) created their static count-based embeddings from the Google n-grams corpus, containing 25 and 28 million of bigrams for the 1960s and for the 1990s correspondingly. This is about two times larger than the COHA sub-corpora for the same time periods. Since the training corpora are different, one can make only preliminary observations concerning the comparative performance of the aforementioned approaches on the GEMS test set. As Frermann and Lapata (2016) put it, 'the use of a different underlying corpus unavoidably influences the obtained semantic representations'.

However, the shared task baselines, 'semantic density' by Sagi et al. (2009), and *word2vec* cosine similarity are directly comparable to ELMo embeddings trained on the test COHA corpora (either as a single model or as two different incremental models). Overall, the SemEval-2020 Task 1 test sets and baseline methods (Schlechtweg, McGillivray, et al., 2020) provide an excellent evaluation test-bed, due to their consistency and multilingual nature. Tables 6.5, 6.6, 6.7 and 6.8 show the performance of all our methods in comparison with the baselines for English, German, Latin and Swedish test sets correspondingly.

The average scores across all the four languages of the SemEval-2020 semantic change shared task for each of the tested configurations are given in Table 6.9. It shows that no single method achieves statistically significant correlation on *all 4* languages, which attests both to the difficulty of the task and the diversity of the test sets. Cosine similarity between Procrustes-aligned CBOW embeddings

Model	**Scores**		
	Baselines		
(Sagi et al., 2009)	-0.186		
Frequency difference (FD)	-0.217		
Count-based (CNT+CI+CD)	0.022		
	Word2vec cosine similarity		
Incremental models	0.210		
Procrustes-aligned models	0.285		
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-train.	0.253 (0.107)	0.198 (0.145)	0.191 (0.121)
+ fine-tuning	0.225 (0.156)	0.124 (0.188)	0.162 (0.199)
ELMo Pre-train.	0.209	0.190	–
+ fine-tuning	0.254	0.220	–
Trained on test	0.138	0.132	–
Incremental	0.076	0.053	–
	Pairwise distance (APD)		
BERT Pre-train.	0.315 (0.078)	0.144 (-0.144)	0.137 (0.104)
+ fine-tuning	0.546* (0.558*)	0.215 (0.362*)	0.368* (0.463*)
ELMo Pre-train.	0.203	0.064	–
+ fine-tuning	**0.605***	0.602*	–
Trained on test	0.291	0.333*	–
Incremental	0.377*	0.302	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-train.	0.175 (-0.018)	0.091 (0.039)	0.009 (-0.026)
+ fine-tuning	0.261 (0.100)	0.170 (0.069)	-0.026 (0.184)
ELMo Pre-train.	0.228	0.235	–
+ fine-tuning	0.223	0.200	–
Trained on test	0.117	0.107	–
Incremental	-0.202	-0.114	–
	Diversity (DIV)		
BERT Pre-train.	0.125 (0.140)	-0.065 (0.243)	-0.106 (0.118)
+ fine-tuning	0.180 (0.008)	-0.127 (-0.165)	0.099 (0.027)
ELMo Pre-train.	0.089	0.072	–
+ fine-tuning	-0.110	-0.064	–
Trained on test	0.161	0.190	–
Incremental	0.462*	0.403*	–

Table 6.5: Results for SemEval-2020 Task 1 sub-task 2 on English: Spearman correlation. Scores for BERT-Large are given in parentheses.

is a very strong approach, consistently outperforming the baselines.[18] Only PRT and APD contextualized methods obtain higher average scores, with fine-tuned ELMo models performing better than fine-tuned BERT. DIV and JSD methods did not manage to outperform the static baseline. For DIV, the reason for this can be that it failed to properly handle cases when a single-sense word has changed its meaning: its diversity can well stay more or less the same, failing to indicate a semantic shift. For JSD, imperfect Affinity Propagation can be blamed, since we saw already in Section 6.3 that it failed to estimate the number of word senses, approximating word frequency instead. It would be interesting to study the errors of DIV and JSD in more detail in the future (considering that in theory these methods are more powerful than PRT and APD), but for now we will focus on PRT and APD.

6.4.4 Closer inspection: contextualized methods are better

Judging only from the average correlation scores, contextualized embeddings do not seem to outshine their static counterparts, especially considering that both ELMo and BERT are more computationally demanding than CBOW. However, closer analysis of per-language results shows that in fact the contextualized approaches outperform the CBOW Procrustes-aligned embeddings by a large margin for *each* of the shared task test sets. Table 6.10 gathers and repeats these per-language results for convenience. It features the scores obtained by our contextualized methods that were the best in most cases:[19] PRT and APD with top layer embeddings from fine-tuned ELMo and BERT. We also again report their performance on the GEMS test set, and the average performance over all *5* test sets (this is our aggregated evaluation score we deem to be the final one).

As can be seen from Table 6.10, different configurations are preferred by different test sets: APD works best on the English and Swedish sets, while PRT yields the best scores for German and Latin. This is why the respective *average* scores of these methods across all test sets are lower and hide their real performance. Admittedly, robustness across test sets (3 out of 4) is an important benefit of the Procrustes-aligned static embeddings approach. However, with the right choice of APD or PRT, contextualized embeddings can improve Spearman correlation coefficients by up to 50%.

The discrepancy between the averaged and the per-language results can be explained by some differences in the test sets. This is not some language-specific property: the English GEMS test set *does not* behave like the English test set from the shared task (i.e., does not clearly prefer APD). In fact, one can observe three interesting groups of test sets with regards to the method they favor and the distribution of gold scores: that is, how uniformly are the degrees

[18] Note that with the shared task test sets, incrementally trained static embeddings consistently perform much worse than their Procrustes-aligned counterparts. The situation is exactly opposite with the GEMS test set, which attests to the importance of the statistical properties of the test set, see subsection 6.4.5 below.

[19] Note that we did occasionally get higher scores for some datasets using other configurations, but rarely and inconsistently.

Model	**Scores**		
	Baselines		
(Sagi et al., 2009)	-0.062		
Frequency difference (FD)	0.014		
Count-based (CNT+CI+CD)	0.216		
	Word2vec cosine similarity		
Incremental models	0.145		
Procrustes-aligned models	0.439*		
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-train.	0.311*	0.154	0.227
+ fine-tuning	0.590*	0.459*	0.463*
ELMo Pre-train.	0.664*	0.616*	–
+ fine-tuning	**0.740***	0.713*	–
Trained on test	0.695*	0.645*	–
Incremental	0.260	0.251	–
	Pairwise distance (APD)		
BERT Pre-train.	-0.003	0.172	0.037
+ fine-tuning	0.427*	0.332*	0.316*
ELMo Pre-train.	0.422*	0.283	–
+ fine-tuning	0.560*	0.482*	–
Trained on test	0.505*	0.397*	–
Incremental	-0.309*	-0.418*	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-train.	0.214	0.121	0.179
+ fine-tuning	0.240	0.356*	0.366*
ELMo Pre-train.	0.417*	0.313*	–
+ fine-tuning	0.434*	0.297*	–
Trained on test	0.257	0.387*	–
Incremental	-0.226	-0.129	–
	Diversity (DIV)		
BERT Pre-train.	0.184	-0.012	0.073
+ fine-tuning	-0.066	-0.089	-0.196
ELMo Pre-train.	0.161	0.138	–
+ fine-tuning	0.291*	0.259	–
Trained on test	0.212	0.118	–
Incremental	0.354*	0.369*	–

Table 6.6: Results for SemEval-2020 Task 1 sub-task 2 on German: Spearman correlation.

Model	**Scores**		
	Baselines		
(Sagi et al., 2009)	0.153		
Frequency difference (FD)	0.020		
Count-based (CNT+CI+CD)	0.359*		
	Word2vec cosine similarity		
Incremental models	0.217		
Procrustes-aligned models	0.387*		
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-train.	0.373*	0.304	0.297
+ fine-tuning	**0.561***	0.420*	0.498*
ELMo Pre-train.	0.414*	0.399*	–
+ fine-tuning	0.360*	0.357*	–
Trained on test	0.370*	0.327*	–
Incremental	0.349*	0.206	–
	Pairwise distance (APD)		
BERT Pre-train.	0.408*	0.235	0.312
+ fine-tuning	0.372*	0.199	0.296
ELMo Pre-train.	-0.015	-0.115	
+ fine-tuning	-0.113	-0.071	–
Trained on test	0.078	-0.117	–
Incremental	0.268	0.143	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-train.	0.461*	0.299	0.416*
+ fine-tuning	0.494*	0.390*	0.429*
ELMo Pre-train.	0.189	0.094	–
+ fine-tuning	0.302	0.154	–
Trained on test	0.474*	0.236	–
Incremental	0.257	0.022	–
	Diversity (DIV)		
BERT Pre-train.	0.203	0.227	0.209
+ fine-tuning	0.045	0.263	0.065
ELMo Pre-train.	0.238	0.260	–
+ fine-tuning	-0.012	0.169	–
Trained on test	0.064	-0.113	–
Incremental	-0.318*	-0.219	–

Table 6.7: Results for SemEval-2020 Task 1 sub-task 2 on Latin: Spearman correlation.

Model	**Scores**		
	Baselines		
(Sagi et al., 2009)	-0.144		
Frequency difference (FD)	-0.15		
Count-based (CNT+CI+CD)	-0.022		
	Word2vec cosine similarity		
Incremental models	-0.012		
Procrustes-aligned models	0.458*		
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-train.	0.261	0.253	0.254
+ fine-tuning	0.185	0.301	0.266
ELMo Pre-train.	0.212	0.169	–
+ fine-tuning	0.252	0.266	–
Trained on test	0.278	0.264	–
Incremental	-0.230	-0.003	–
	Pairwise distance (APD)		
BERT Pre-train.	0.415*	0.390*	0.402*
+ fine-tuning	0.254	0.375*	0.163
ELMo Pre-train.	0.573*	0.457*	–
+ fine-tuning	0.569*	**0.610***	–
Trained on test	0.479*	0.566*	–
Incremental	0.169	-0.031	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-train.	-0.048	-0.005	0.040
+ fine-tuning	-0.120	-0.047	-0.083
ELMo Pre-train.	0.197	0.052	–
+ fine-tuning	-0.146	-0.065	–
Trained on test	0.051	-0.079	–
Incremental	0.022	0.184	–
	Diversity (DIV)		
BERT Pre-train.	0.083	-0.117	-0.037
+ fine-tuning	0.081	-0.056	-0.116
ELMo Pre-train.	0.335	0.425*	–
+ fine-tuning	0.150	0.165	–
Trained on test	0.071	0.104	–
Incremental	0.343	0.252	–

Table 6.8: Results for SemEval-2020 Task 1 sub-task 2 on Swedish: Spearman correlation.

of semantic change distributed in the gold data. For example, this distribution can be skewed to the left (closer to zero), meaning that most words in the test set have not changed, or to the right (closer to 1), meaning that most words in the test set have changed significantly. We explain this measure in more detail in subsection 6.4.5. The test set groups are as follows:

1. group 1 (Latin and German from SemEval-2020 Task 1) exhibits rather uniform gold score distributions and prefers PRT;
2. group 2 (English and Swedish from SemEval-2020 Task 1) is characterized by more skewed gold score distributions and prefers APD;
3. group 3 (GEMS) is in between, with no clear preference.

Interestingly, the method which produces a more uniform predicted score distribution (APD) works better for the test sets with skewed gold distributions, and the method which produces a more skewed predicted score distribution (PRT) works better for the uniformly distributed test sets. Furthermore, there is a strong negative Spearman rank correlation between the median gold score of a test set and the performance of the APD method with fine-tuned ELMo models on this test set; again, see subsection 6.4.5 for further discussion on this point.

Simply averaging the PRT and APD estimations and using them as final predictions yields surprisingly robust results (see the 'PRT/APD' rows in Table 6.10). For individual test sets, the performance of this approach usually lies in between PRT and APD, but when averaged over all five test sets, it ranks higher than any individual approach, and this effect holds for both ELMo and BERT, although the highest score is observed for the former. Thus, it seems that the APD and PRT methods are indeed complimentary and together act as a top-performing ensemble of the models, with the additional benefit of not having to worry about what method to choose.

Table 6.10 also supports the previous observation that ELMo-based models perform better than BERT for lexical semantic change detection (at least in the ranking sub-task). The only test set for which this is not the case is Latin,[20] while on GEMS, ELMo and BERT are approximately on par. With the ensemble approach ('PRT/APD') ELMo embeddings are also on average better than BERT embeddings. One possible explanation is that our ELMo models were pre-trained on lemmatized Wikipedia corpora and thus better fit the lemmatized historical corpora. The BERT models were pre-trained on raw corpora, and fine-tuning them on lemmatized data proves less successful. This is of course not an advantage of the ELMo architecture *per se*; however, easy and fast training from scratch on the respective Wikipedia corpora for each test set was possible only because of much lower computational requirements of ELMo compared to

[20] The SemEval-2020 Latin test sets and corpora are indeed peculiar: 1) homonyms in them are followed by '#' and the sense identifier, which is of course not the case for Latin Wikipedia, on which our contextualized models were pre-trained 2) the sizes of the C_1 and C_2 corpora are very imbalanced, with the latter being four times larger than the former.

Model	**Average Spearman correlation**		
Baselines	**(Sagi et al., 2009)**		-0.060
	Frequency (FD)		-0.083
	Count (CNT+CI+CD)		0.144†
CBOW cosine distance	**Incremental**		0.140
	Procrustes		0.392†††
Contextualized embeddings	**Top layer**	**All layers**	**Top 4 layers**
	Cosine similarity (PRT)		
BERT Pre-trained	0.278††	0.233	0.229
+ fine-tuning	0.373††	0.320††	0.338††
ELMo Pre-trained	0.375††	0.344††	–
+ fine-tuning	0.402††	0.389††	–
Trained on test	0.370††	0.342††	–
Incremental	0.114†	0.127	–
	Pairwise distance (APD)		
BERT Pre-trained	0.237††	0.163†	0.203†
+ fine-tuning	0.363†††	0.241††	0.297†
ELMo Pre-trained	0.296††	0.172†	–
+ fine-tuning	0.405†††	**0.406**†††	–
Trained on test	0.338††	0.295†††	–
Incremental	0.126††	-0.001†	–
	Jensen-Shannon divergence (JSD)		
BERT Pre-trained	0.181†	0.125	0.203†
+ fine-tuning	0.176†	0.223††	0.186††
ELMo Pre-trained	0.251†	0.196†	–
+ fine-tuning	0.197†	0.156†	–
Trained on test	0.225†	0.163†	–
Incremental	-0.037	-0.009	–
	Diversity (DIV)		
BERT pre-trained	0.154	-0.028	0.065
+ fine-tuning	0.036	0.010	-0.042
ELMo pre-trained	0.206	0.224†	–
+ fine-tuning	0.080†	0.132	–
Trained on test	0.127	0.075	–
Incremental	0.210†††	0.201††	–

Table 6.9: Correlation scores for the methods under analysis on SemEval-2020 Task 1 Subtask 2 averaged over four languages. The number of † denotes the number of languages for which the correlation was statistically significant ($p < 0.05$).

Method	*English*	*German*	*Latin*	*Swedish*	*GEMS*	*Average*
SemEval-2020 Task 1 baselines						
FD	-0.217	0.014	0.020	-0.150	0.068	0.094
CNT+CI+CD	0.022	0.216	0.359*	-0.022	0.256*	0.166
Word2vec CBOW cosine distance						
Incremental	0.210	0.145	0.217	-0.012	**0.424***	0.197
Procrustes-aligned	0.285	0.439*	0.387*	0.458*	0.235*	0.361
Fine-tuned contextualized embeddings (top layer)						
ELMo PRT	0.254	**0.740***	0.360*	0.252	0.323*	0.386
ELMo APD	**0.605***	0.560*	-0.113	**0.569***	0.323*	0.389
ELMo PRT/APD	0.546*	0.678*	0.036	0.546*	0.360*	**0.433**
BERT PRT	0.225	0.590*	**0.561***	0.185	0.394*	0.391
BERT APD	0.546*	0.427*	0.372*	0.254	0.243*	0.368
BERT PRT/APD	0.498*	0.537*	0.431*	0.267	0.332*	0.413

Table 6.10: Spearman correlations per test set for our best methods. * denotes statistically significant correlation ($p < 0.05$).

BERT (on average, our ELMo models had two times less parameters than the BERT models).

Note also that using BERT-*Large* instead of BERT-*Base* does not seem to improve BERT results much. We replicated our experiments with BERT-*Large* for the English SemEval-2020 Task 1 test set. The corresponding results are given in parentheses in Table 6.5. They follow the trends outlined above, with the APD method being the best. Using the *Large* model does yield slightly higher correlation than using the *Base* model (0.558 versus 0.546 with the fine-tuned model and top layer embeddings), but it is still lower than the corresponding ELMo results (0.605). Considering the marginal value of these improvements and even higher computational requirements of BERT-*Large*, we did not test it with the other test sets, leaving this for future work (although we expect that the outcome will be similar).

In the post-evaluation phase of the shared task, we submitted predictions obtained with the optimal system configurations: fine-tuned ELMo + APD for English and Swedish, fine-tuned ELMo + PRT for German, and fine-tuned BERT + PRT for Latin. It reached the average Spearman correlation of 0.618 and, at the time of writing, it is the best publicly available post-evaluation submission for SemEval-2020 Task 1 Subtask 2 (‘UiO-UVA’ team). Certainly, this was made possible only because we were able to analyze the statistical properties of the test sets (which were hidden in the evaluation phase) and relate them to different semantic change estimation methods. However, we emphasize that this is not the same as training on a test set: the employed methods still did not know anything about the gold annotations and took only historical corpora and target word lists (without any scores) as their inputs. Note also that using the ensemble 'PRT/APD' method avoids the need to know the gold score distribution beforehand and can be used as is on any test set, outperforming the non-contextual baselines in most cases.

6.4.5 Dependency between score distribution in the test set and method preference

In this subsection, we provide some additional analysis of the statistical properties of the test sets which influence their 'preferred' contextualized semantic change detection methods.

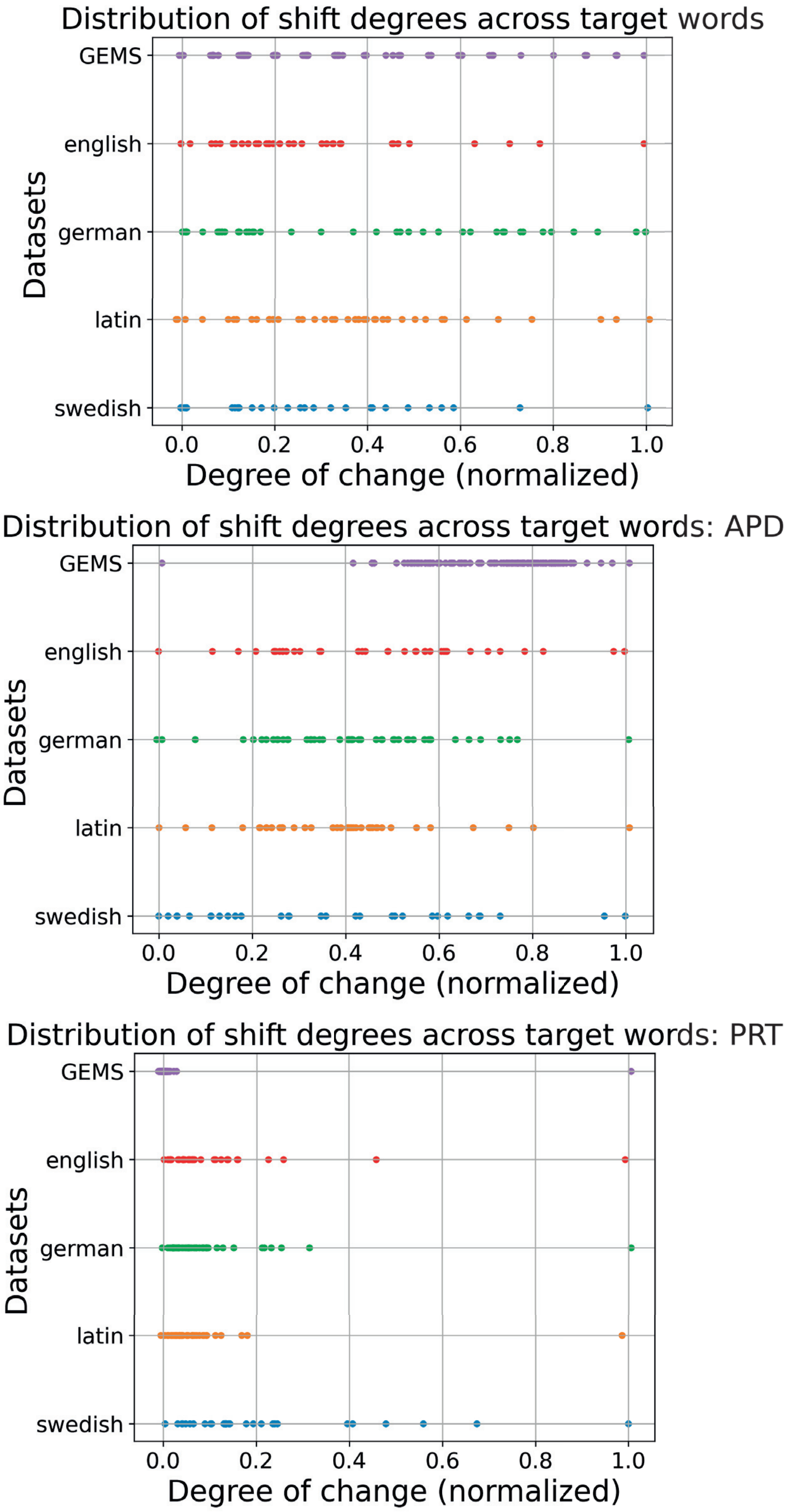
In the top part of Figure 6.8, we show how different the five semantic change test sets are in terms of how the *gold* scores are distributed across them. It is clearly visible on the plot that in some test sets, the normalized gold scores are skewed to the left, while some have a more uniform distribution. The middle and bottom parts of Figure 6.8 show the distributions of the *predicted* scores produced by the APD and PRT methods (with fine-tuned ELMo embeddings). PRT tends to squeeze the majority of predictions near the lower boundary (no semantic change), with a low median score. On the opposite, APD distributes its predictions in a much more uniform way, with a higher median score. Counter-intuitively, skewed gold distributions favor uniform predictions and vice versa.

The grouping differences can be quantified with respect to the median gold score (after unit-normalization). Figure 6.9 shows the dependency of the PRT and APD performance on the median score of the gold test set. The dots here are the performance values of PRT or APD methods on different test sets. English and Swedish test sets are in the left part of the plot with the median gold scores of 0.200 and 0.203 correspondingly. German, GEMS and Latin are on the right with 0.266, 0.267 and 0.364 correspondingly. There is a perfect negative Spearman rank correlation ($\rho = -1$) between the median gold scores of these five test sets and the performance of APD semantic change detection method on each of them (with fine-tuned ELMo embeddings). We currently do not have a plausible explanation for this behavior. Admittedly, the number of data points is not large (5), and the effect can be spurious, but this is doubtful, considering that it is manifested across two independent neural architectures.

6.4.6 Empirical evaluation summary

To sum up, our main findings from the empirical evaluation of contextualized methods for lexical semantic change estimation (ranking sub-task) are as follows:

1. Contextualized embeddings outperform the methods based on static embeddings (and other distributional baselines) in all the five test sets we used.



Distribution of shift degrees across target words

Distribution of shift degrees across target words: APD

Distribution of shift degrees across target words: PRT

Figure 6.8: Top: distribution of semantic change degree in the **gold** data; middle: distribution of scores predicted by the **APD** method; bottom: distribution of scores predicted by the **PRT** method.

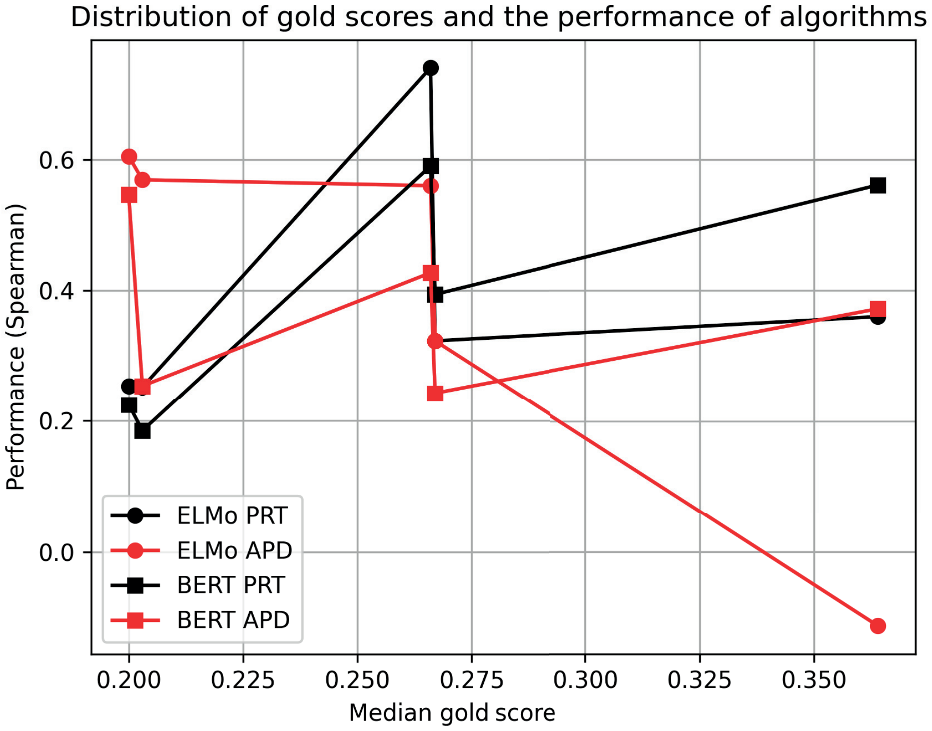


Figure 6.9: Performance of the PRT and APD contextualized methods depending on the median gold score.

2. In three out of five test sets, ELMo consistently outperforms BERT, while having much less parameters and being much faster in training and inference.

3. Inverted cosine similarity of averaged contextualized token embeddings (PRT) and average pairwise distance between these embeddings (APD) are the two best-performing change detection methods. The methods based on token diversity calculation (DIV) or on clustering (JSD) turned out to be inferior. One of the reasons for that can be explained by a quote from Schlechtweg, Hätty, et al. (2019): 'dispersion measures are strongly influenced by frequency and very sensitive to different corpus sizes'. Future work will show whether controlling for these factors can improve the results for DIV and JSD (which in theory should be very powerful methods).

4. Different test sets show preference to either PRT or APD method. This preference is strongly correlated with the distribution of gold scores in a test set.

5. While it may indicate that there are biases in the available test sets,

this finding remains yet unexplained. We did not manage to find any other property of the test sets (word frequencies, the width of the time gap between the historical corpora, etc) which would correlate with the performance of either PRT or APD method.

6. In the realistic case of not being able to find out the gold scores distribution beforehand, it is recommended to use the average of the PRT and APD predictions (‘PRT/APD’), which proved to be a very robust ‘ensemble’ approach.

6.5 Qualitative analysis

In this section, we qualitatively examine the output of the contextualized methods evaluated in the previous section. We analyze both the examples of undoubtedly real semantic shifts and the examples of controversial nature. The latter examples are arguably more important for future studies, and we propose a working categorization of such controversial cases.

6.5.1 Good examples of detected shifts

For many words, the scores produced by our semantic change modeling methods do signal a new emergent sense. As an example, let us consider the word ‘*cell*’. The dataset from Tsakalidis et al. (2019) (based on the Oxford English Dictionary definitions) mentions this word as having acquired a new meaning of ‘MOBILE PHONE’ after the year 2000. We will look at the average of the PRT and APD change scores calculated for this word using its contextualized token embeddings inferred from consecutive pairs of COHA decades (1960s-1970s, 1970s-1980s, 1980s-1990s, 1990s-2000s). As a contextualizer, we employ a single ELMo model trained on the whole COHA corpus.

Recall that PRT and APD (and their average as well) produce as an output a measure of how strong the semantic change of a query word was between two time bins; this measure characterizes a pair of decades in our case. The ‘*cell*’ experienced a change of 0.6727 in the 1970s compared to the 1960s[21] (arguably corresponding to the start of its widespread usage in the biological sense). After that, the change degrees were smaller, with 0.6694 in the 1980s and 0.6718 in the 1990s. However, the 2000s saw the change degree of 0.6950 compared to the 1990s (the highest change for this word across all decades), most likely reflecting the new ‘MOBILE PHONE’ sense.

Unlike the ‘static’ word embedding approaches described in the previous chapters, using contextualized embeddings allows us to visually explore the sentences where a given word is used in different senses, according to our model. For this purpose, we use Principal Component Analysis (PCA) to reduce the 1024-dimensional ELMo representations of each ‘*cell*’ occurrence in our training

[21] The average change degree for the words from the same frequency tier in the 1960-1970s decade pair is 0.699, with the standard deviation of 0.103, so in terms of absolute value this is not actually high. See more on this in the next subsection 6.5.2.

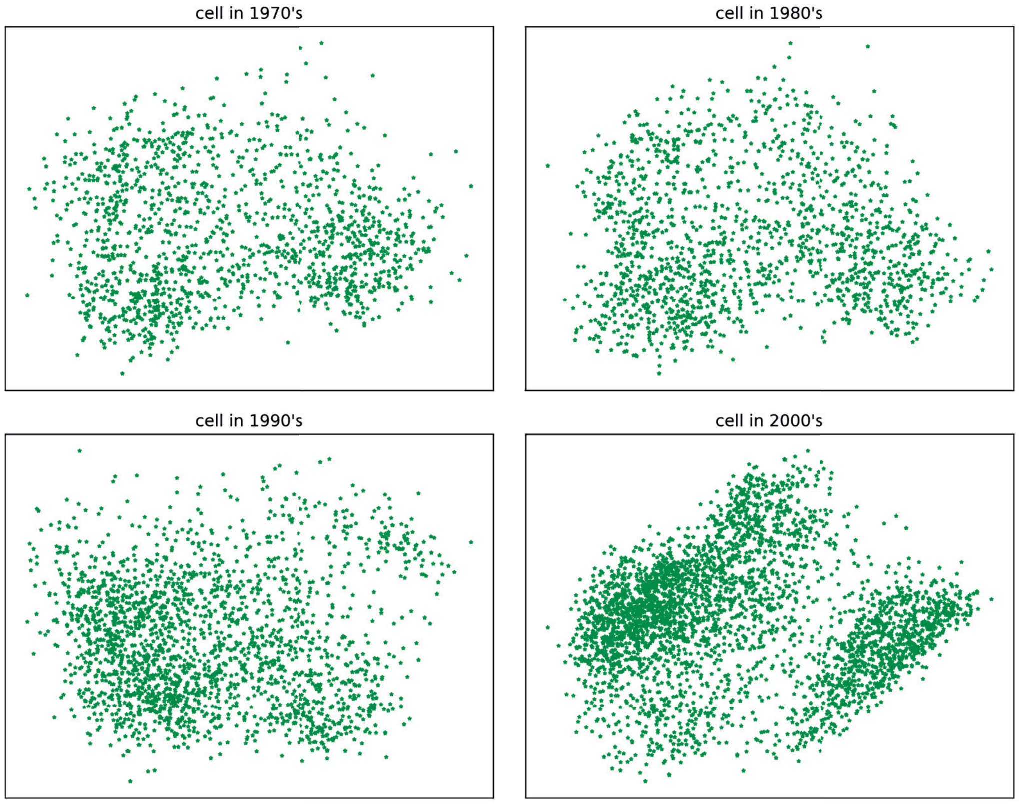


Figure 6.10: PCA projections of ELMo representations of each occurrence of the word '*cell*' in four different decades: actual semantic shift.

corpora to flat 2-dimensional projections. Figure 6.10 shows these projections for the decades of 1970s, 1980s, 1990s and 2000s. Note that PCA is a lossy dimensionality reduction technique, and the resulting visualizations only roughly reflect the similarities and dissimilarities of various '*cell*' occurrences in the original high-dimensional vector space. However, one can still make some observations.

Even at a glance, it is possible to see that the 2000s semantic change is caused by radical changes in the groupings of the '*cell*' token embeddings. The three previous decades are all characterized by a rather vague separation of this word's usages into two clusters (at the left and at the right part of the vector space). In the 2000s, we observe the appearance of a new cluster: now there are two strong clusters to the left and a third one to the right. But what senses do these clusters correspond to? Fortunately, since each point on the plot represents a particular '*cell*' occurrence from a particular decade's sub-corpus, we can retrieve these occurrences from the texts and analyze them qualitatively. In this way, we observe that in the 1970s, 1980s and 1990s, the right-hand cluster mostly contains sentences with '*cell*' in the sense of 'PRISON CELL'. We give some examples in 5, where each sentence is from a different decade:

(5)

1. 'I'd known Archie Meltzer, the chief turnkey on duty, for over ten years, but you wouldn't have known it from the way he processed me for the **cells**.'
2. 'It also happened to me in a jail **cell**, Peb.'
3. 'If she had been writing to somebody in the darkness of her prison **cell**, what had she done with the message?'

As it turns out, the left cluster (consistently increasing its relative size over time) mostly contains sentences with '*cell*' in the biological sense, with examples given in 6:

(6)

1. 'The sexual **cells** of Pyronema show this in ascomycetes.'
2. 'It's how a **cell** decides whether it becomes a muscle **cell** or a skin **cell**.'
3. 'If those **cells** are found to be cancerous after being sent to a lab, that's a definite diagnosis.'

After exploring the points in the 2000s plot in the same way, one observes that the two clusters on the left correspond to the old senses of '*cell*' (biological still at the bottom and prison at the top). But the new large cluster on the right almost exclusively consists of sentences mentioning '*cell*' in the sense of 'MOBILE PHONE' (see examples in 7 and Figure 6.11 displaying these clusters with labels).

(7)

1. 'But how well do the service providers fulfill that objective, and what about the other health and safety risks - exposure to radio waves and potentially fatal driver distraction - that the growing use of **cell** phones raise?'
2. 'Gilles swatted Adriana on the upper arm as he walked past, nearly dislodging the **cell** phone she had balanced between her chin and her left shoulder.'
3. 'You still have the same **cell** number.'

Interestingly, the 'MOBILE PHONE' cluster has already started to appear in the 1990s (the small group of occurrences at the top right corner of the plot). Examples of sentences from this cluster are given in 8 below. However, it was too small (not more than 50 tokens out of a total of 2160 '*cell*' occurrences in the 1990s corpus), and thus did not cause a high change score. Only in the 2000s did the number of usages for this sense become large enough to achieve the highest observed change score for this word (and create a clearly visible separate cluster of vector token representations).

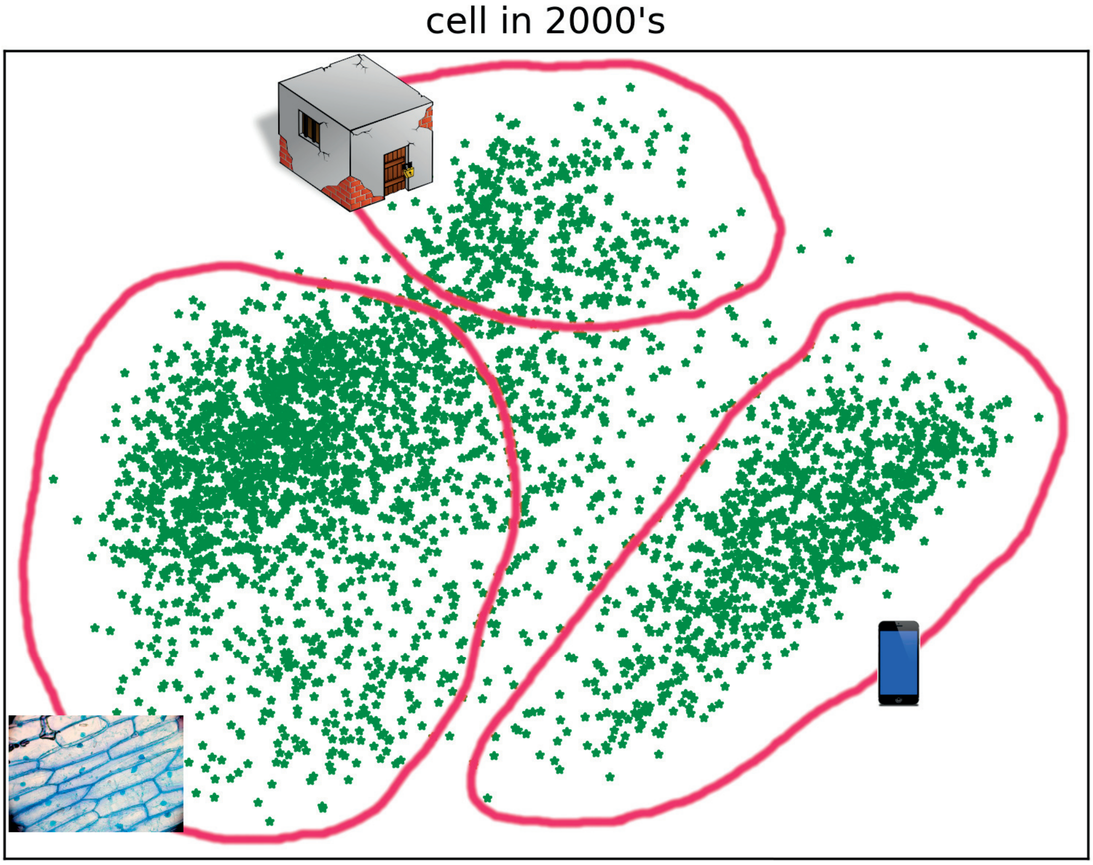


Figure 6.11: PCA projection of ELMo token representations of each occurrence of the word ‘*cell*’ in the 2000s, with clusters labeled with senses.

(8)

1. ‘Congressman John Boehner joined in by **cell** phone from Florida.’
2. ‘A lot of people seem to forget their own **cell** phone numbers.’
3. ‘Just use your **cell** phone.’

One can also visualize ELMo token embeddings for ‘*cell*’ across all five time bins, as shown in Figure 6.12. Here, PCA dimensionality reduction is performed for all occurrences of this word (about 7 500 total), and thus we see how usages from different decades are grouped in relation to each other. It is clear that the top right cluster is inhabited almost exclusively with the occurrences from 2000s and to a less extent the 1990s. Not surprisingly, it contains sentences where ‘*cell*’ is used in the ‘MOBILE PHONE’ sense. At the same time, in other parts of the plot, occurrences from all decades are dispersed more or less uniformly, supporting our previous observation that in the 60s, 70s and 80s, this word did not experience significant semantic shifts.

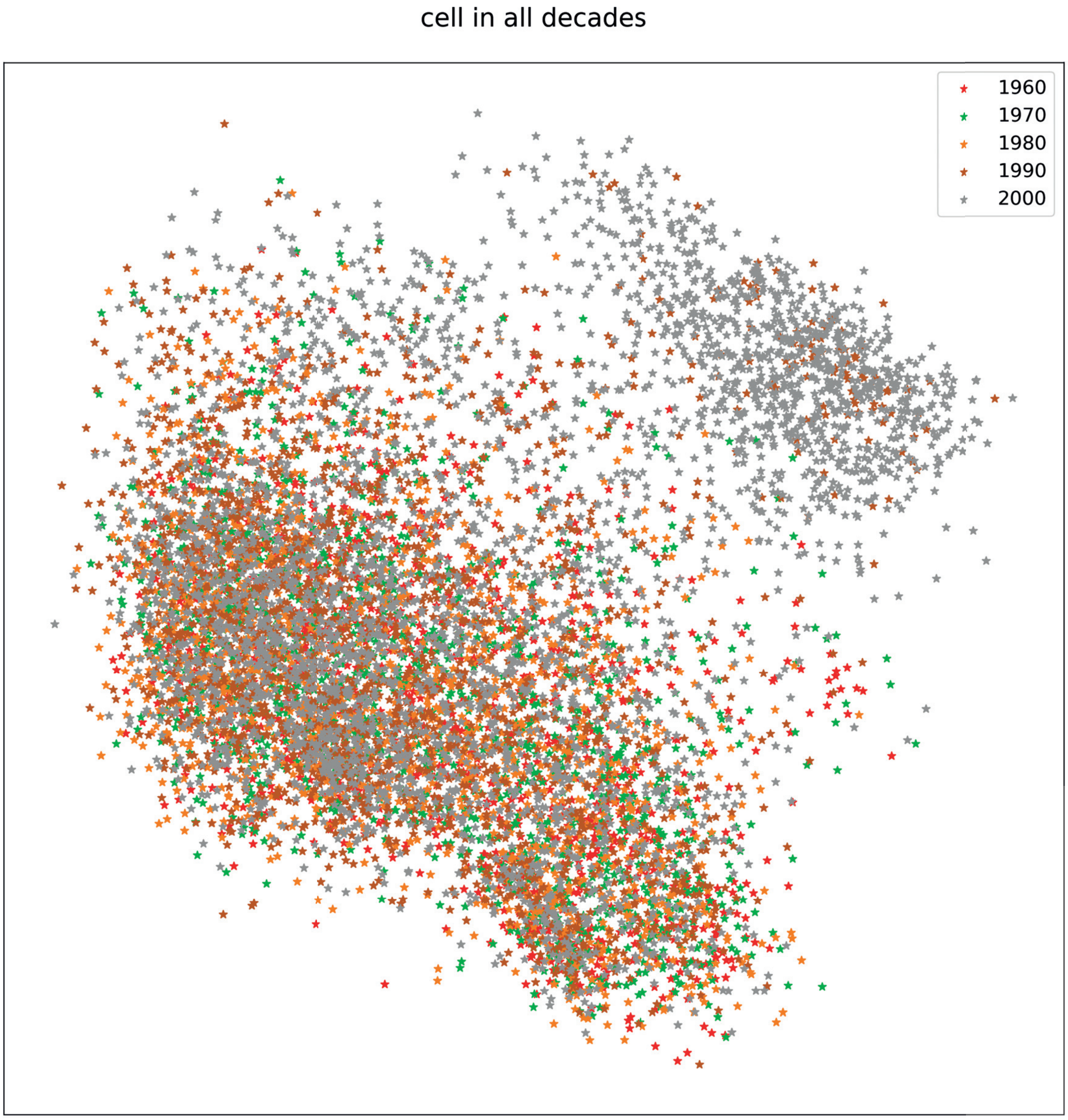


Figure 6.12: PCA projection of ELMo contextualized representations of occurrences of the word '*cell*' in all five COHA decades (single COHA model).

In the case of '*cell*', the groupings of ELMo representations and the changes detected by our methods are undoubtedly connected to a new sense emerging (thus, a diachronic semantic shift). The relations between different senses of '*cell*' fall into the category of *homonymy*, where word senses are not directly related to each other (at least, synchronically). However, one can trace the cases of *polysemy* as well, where senses are synchronically related to each other. As an example, let us look at the adjective '*virtual*'. In the COHA corpus, it was always quite fluid in its meaning (as measured by the single ELMo model and the PRT/APD method), but it experienced the strongest change in the 1990: 0.7692 (see its scores for other decades in Table 6.11).

Before 1990s, '*virtual*' was used mostly in two closely related senses: 'BEING SUCH IN ESSENCE OR EFFECT THOUGH NOT FORMALLY RECOGNIZED OR ADMITTED' (major one) and 'RELATED TO A HYPOTHETICAL PARTICLE WHOSE EXISTENCE IS INFERRED FROM INDIRECT EVIDENCE' (minor).[22] However, the 1990s saw the emergence of a large number of '*virtual*' usages in the sense of 'SIMULATED ON A COMPUTER OR COMPUTER NETWORK', especially in the expression 'virtual reality' (almost one third of all usages). This sense is related to the previous ones, thus manifesting a case of a polysemous word. The emergence of a new related sense in the 1990s is captured by contextualized embedding based methods, producing a higher change score for this time bin in comparison to the previous 1980s decade. We can also observe a much weaker change score in the 2000s, which is supported by the manual inspection of the occurrences showing that in the 2000s, '*virtual*' was still used a lot in this third sense (although, interestingly the 'virtual reality' expression itself almost came out of usage, constituting now only 6% of all '*virtual*' occurrences).

Figure 6.13 plots the ELMo token embeddings for all occurrences of '*virtual*' across five COHA decades. The 'SIMULATED ON A COMPUTER OR COMPUTER NETWORK' usages occupy the left part of the plot, with the 'virtual reality' phrases concentrated in the left top corner (as confirmed by manual inspection). The left part contains almost exclusively the occurrences from the 1990s and from the 2000s, while the left top corner is dominated by the 1990s.

Note that the 1990s, when the word '*virtual*' changed most, also saw the sharpest increase in its frequency: from six instances per million tokens to 22 instances per million. The PRT/APD scores and the frequency changes (δ) for '*virtual*' and '*cell*' are given in Table 6.11: positive δ represent an increase in corpus frequency, and negative δ represent a decrease in corpus frequency (it can be seen that both words belong to the high-frequency tier, with '*virtual*' being somewhat less frequent). In fact, for '*virtual*', the PRT/APD scores per decade pair are perfectly correlated with the changes in frequency, either absolute or normalized to instances per million: Spearman rank correlation is 1.0, with p-value 0.

However, this does not mean that change detection methods based on contextualized embeddings always simply approximate frequency changes. We

[22] The definitions borrowed from the online version of the Merriam-Webster dictionary, https://www.merriam-webster.com/.

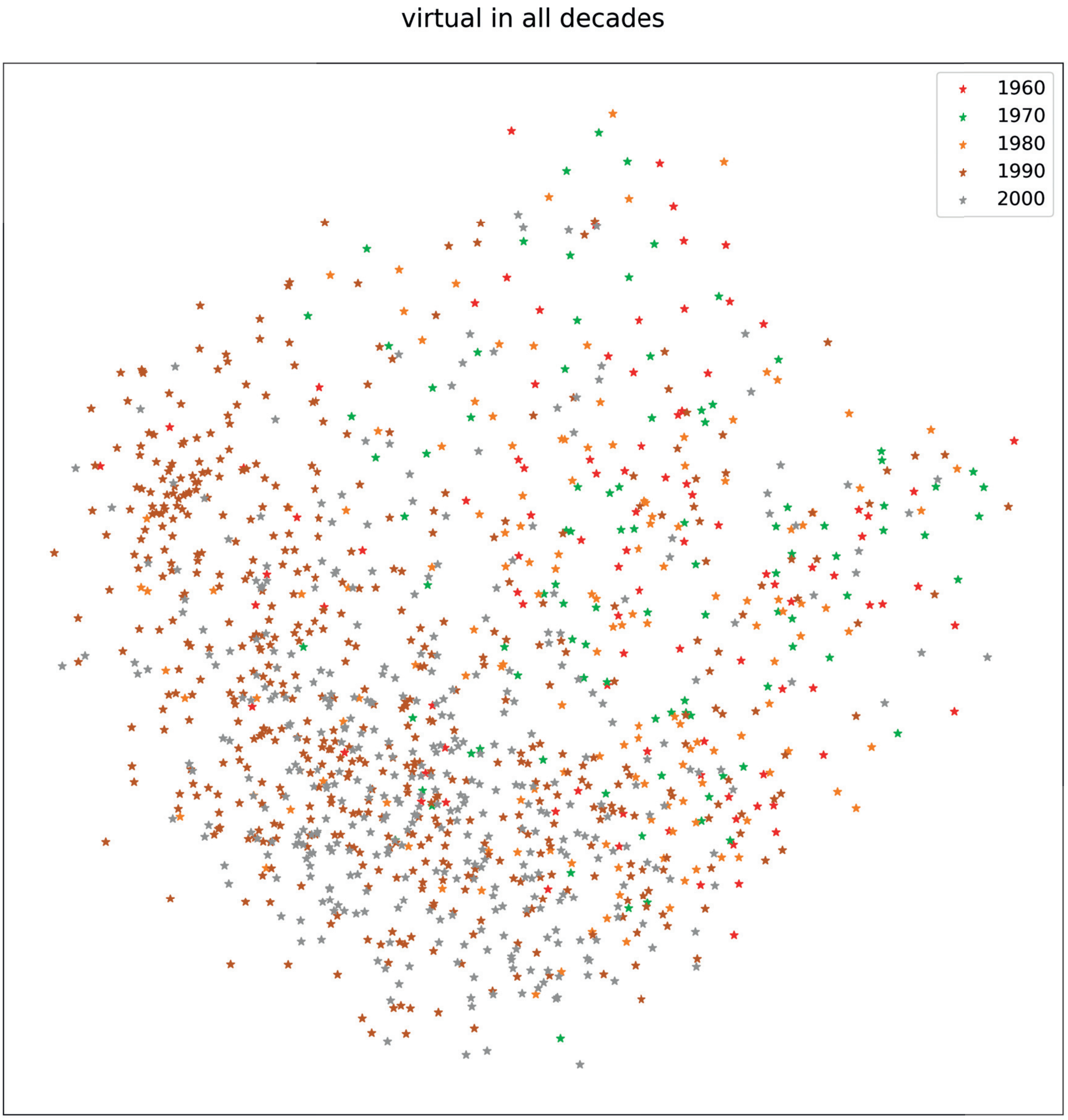


Figure 6.13: PCA projection of ELMo contextualized representations of occurrences of the word ‘*virtual*’ in all five COHA decades (single COHA model).

Decades	PRT/APD		Absolute frequency δ		IPM δ	
	‘*cell*’	‘*virtual*’	‘*cell*’	‘*virtual*’	‘*cell*’	‘*virtual*’
1960s-1970s	0.6727	0.7423	244	-2	11	0
1970s-1980s	0.6694	0.7478	118	50	2	2
1980s-1990s	0.6718	0.7692	597	381	27	16
1990s-2000s	0.6950	0.7401	1 286	-143	52	-6

Table 6.11: PRT/APD predicted semantic change scores and their frequency differences (δ) across five COHA decades for the words ‘*cell*’ and ‘*virtual*’. IPM stands for ‘instances per million’.

already showed how this is not the case for the DIV method in Section 6.3. Table 6.11 demonstrates this for the PRT/APD method as well. Unlike with '*virtual*', for the word '*cell*', its predicted scores do not exactly follow its frequency changes. Although '*cell*' frequency counts increased in the 1990s (compared to the 1980s) much more than in the 1970s (compared to the 1960s), PRT/APD still predicts stronger change for the latter than for the former. Formally speaking, Spearman rank correlation between '*cell*' PRT/APD scores and its frequency changes is only about 0.8, and it is not statistically significant (p-value 0.2). Additionally, this correlation is observed only if we look at the data limited to *one* particular word (for example, '*cell*' or '*virtual*'). If we concatenate the data even for only two words '*cell*' and '*virtual*', no correlation can be found between PRT/APD scores and frequency changes at all. Adding more different words still leaves these data columns uncorrelated. This indicates that in general, PRT/APD scores cannot be predicted from looking at words' frequencies in different time bins.

So far so good: the contextualized embedding-based methods not only demonstrate high scores on the evaluation sets, they also produce interpretable predictions corresponding to well-known diachronic semantic shifts. But let us also look at their darker side.

6.5.2 Controversial examples of detected shifts

Unfortunately, the picture is not as clear if one looks beyond hand-picked examples. As mentioned above, the change score of '*cell*' when comparing the 2000s to the 1990s was 0.6950 . But the absolute values of this score are not very informative themselves. It is not the case, for example, that the scores higher than 0.7 always point at some breaking points in the history of a word's semantics. We can observe much stronger bursts which do not yield to such an explanation.

To illustrate this, we calculated the PRT/APD scores between the five COHA decades for a word list consisting of all words we used in subsection 6.3.2 (i.e., 3 000 lexical entries occurring in all five COHA sub-corpora and representing three frequency tiers), all words from the SemEval-2020 Task 1 English test set and all words from the GEMS test set. After excluding numerals and function words, the size of this list is 2 995 entries. Let us look at the subset consisting of words with a total frequency of more than 1 000 occurrences across all decades (to discard unstable rare words representations).

Table 6.12 lists 10 points of the sharpest change across all words and consecutive decade pairs. None of them can be immediately interpreted in any meaningful way (as acquiring or losing a sense). The question arises: what is the cause of these bursts?

Looking closely at these cases reveals three general classes of words which trigger high semantic change score as measured by PRT/APD and at the same time do not represent proper semantic shifts. The classes are (colors correspond to those in Table 6.12):

Word	Decade pair	Change score	Class
‘*banish*’	1980s-1990s	0.7940	Proper name
‘*designate*’	1980s-1990s	0.7921	Context-dependent
‘*mg*’	1980s-1990s	0.7912	Data burst
‘*progressive*’	1990s-2000s	0.7824	Context-dependent
‘*indirectly*’	1990s-2000s	0.7803	Data burst
‘*form*’	1990s-2000s	0.7801	Context-dependent
‘*subsequently*’	1980s-1990s	0.7800	Context-dependent
‘*neutral*’	1990s-2000s	0.7792	Data burst
‘*traditionally*’	1990s-2000s	0.7791	Syntactic change
‘*pointed*’	1960s-1970s	0.7785	Context-dependent

Table 6.12: 10 points of the strongest change in five decades of COHA (as measured by PRT/APD). Word color indicates its class.

1. Words of very context-dependent meaning (‘*designate*’, ‘*progressive*’, etc): their token embeddings are very different from each other (and thus change scores are high) when compared either synchronically or diachronically.

2. Words frequently used in a very specific context in a particular time bin, different from other periods (‘*mg*’, ‘*indirectly*’, etc). It can be looked at either as a result of (unintended) domain shifting when building a corpus or as real context variance which did not lead to the emergence of a new lexicographic sense (or losing an old one). We will also call such cases ‘data bursts’. There is an interesting sub-type of this class:

    - words used as a proper name in a particular time bin (‘*banish*’, etc.); this leads to extremely high context variance and the emergence of strongly detached token clusters.

3. Words undergoing syntactic changes, not semantic ones; see below.

Figure 6.14 shows the PCA projections of token embeddings for four of the words from Table 6.12 across the whole COHA time span we use. Let us describe these diachronic vector spaces more closely to explain the nature of each controversial word class.

‘*Progressive*’ (in the bottom left part of the plot) belongs to the 1st class and presents the easiest case to explain. As can be seen from the plot, the occurrences from all five decades are spread more or less uniformly over the vector space. There are no regions inhabited by occurrences only from some subset of the decades. This means no sense was acquired or lost at any point in time. The reason for the high absolute value of the change score is the generic meaning of the word itself. Actually, it featured high change scores in all the previous

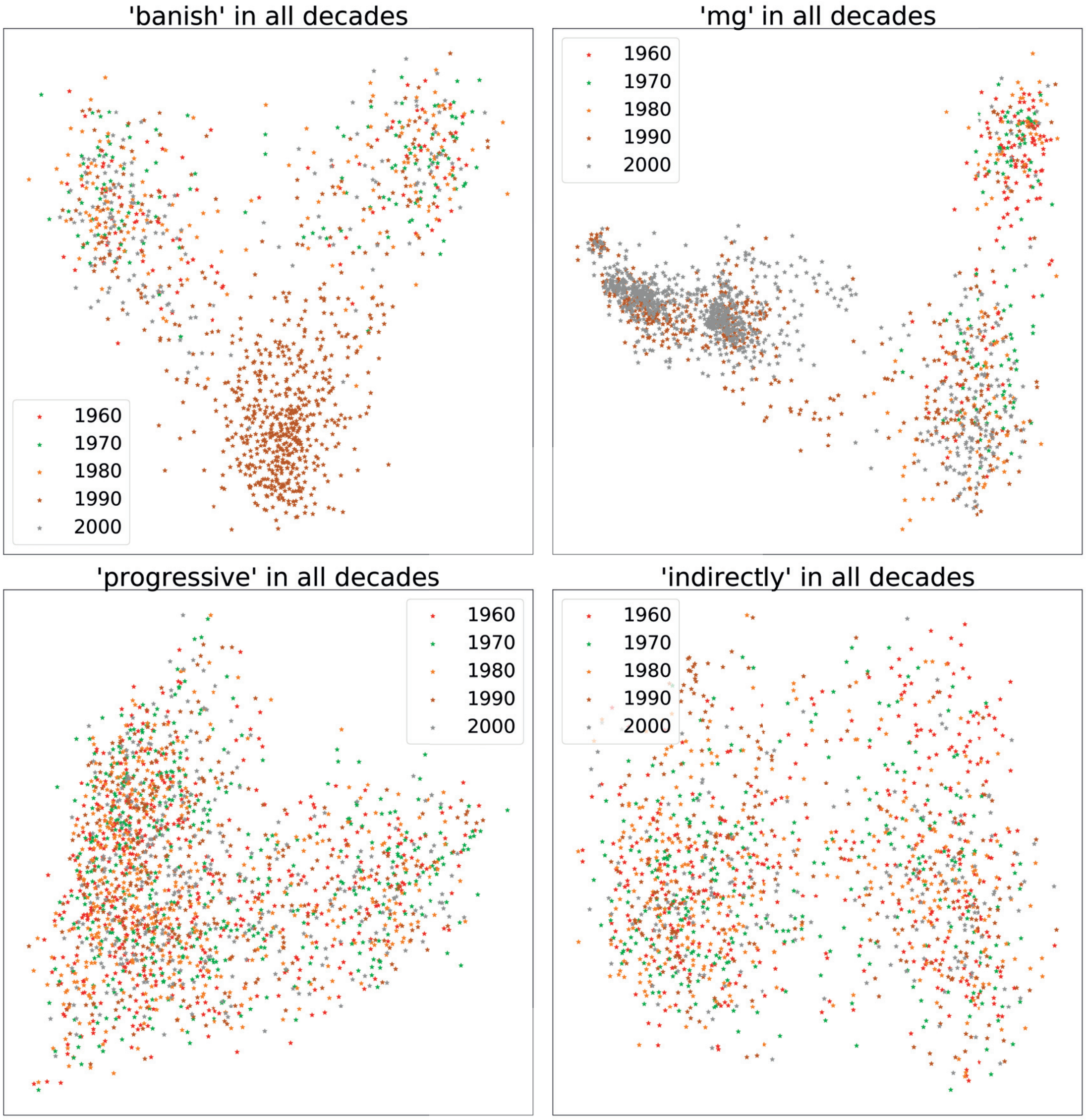


Figure 6.14: PCA projections of ELMo token embeddings for '*banish*', '*mg*', '*progressive*' and '*indirect*' across all five COHA decades.

decade pairs as well: 0.7814, 0.7795, 0.7783. Its contexts are so diverse and 'fluid' that our methods detect strong change whatever corpora are under comparison. In this respect, '*progressive*' (as well as '*designate*', '*form*' and other similar items) behaves much like function words: their contextualized embeddings are in a constant flux. Such cases can be traced and discarded when we have a long sequence of time bins (for example, five decades of COHA) clearly showing the constant character of the changes. However, if looking at one pair of time bins only (like in the previous Section 6.4), a researcher can be mistaken into concluding that an actual semantic shift is going on.

‘*Indirectly*’ and ‘*mg*’ (bottom and top right parts of the plot correspondingly) belong to the 2nd class and they do reflect some actual changes in the text (although possibly not proper semantic shifts). The PCA projection for ‘*indirectly*’ features a small constellation of the 1990s occurrences in the top left corner. Otherwise, the occurrences from different time bins are spread uniformly, so this must be the reason of the detected ‘change’. Indeed, we see that for this word, high change scores are computed both in the 1990s (0.7785) and in the 2000s (0.7803), while before that the change scores were much lower. Accordingly, it seems that something had happened to ‘*indirectly*’ in the 1990s and then went back to ‘normal’ in the 2000s. Manual inspection of the 1990s-specific cluster reveals sentences like those in example 9:

(9)

1. ‘Lane now holds 1,966,692 shares directly and **indirectly**, worth $ 17,700,228.’
2. ‘Parshall now holds 300 Class A shares **indirectly**, worth $ 3,975.’

All of them are excerpts from a long text titled ‘Depressed shares are a hit with bargain-hunting execs Banks, utilities among winners’, apparently published in ‘Insider trading’ magazine in 1994. It abounds with reports on various persons holding various amounts of shares directly or indirectly. This type of texts is quite unusual for the COHA as a whole: there are no sentences mentioning both ‘*hold*’ and ‘*indirectly*’ simultaneously in other decades, except only one such sentence in the 1980s. Meanwhile, the 1990s sub-corpus has 27 of them (that is approximately the size of the outlier cluster we see in the plot). The 2000s sub-corpus does not include such texts any more, and thus we observe an equally strong change back when moving from the 1990s to the 2000s.

For the word ‘*mg*’ (milligram) the situation is similar, except that the change in the 1990s (change score of 0.7921) was the only burst (for other decade pairs, the change scores do not exceed 0.71). It means that something changed in the 1990s and stayed like this through the 2000s as well. Inspecting Figure 6.14 (top right plot) shows that there is indeed a clearly separated cluster consisting only of the 1990s and 2000s tokens. In the corpus, they always occur in the phrase ‘mg cholesterol’, in sentences like ‘Per serving: 525 calories, 34 gm protein, 18 gm carbohydrates, 36 gm fat, 674 **mg cholesterol**, 6 gm saturated fat, 409 mg sodium’, of course being part of dish recipes published in newspapers and magazines. The word ‘*cholesterol*’ has occurred in COHA before the 1990s, but never in a similar context (we observe 128 occurrences of ‘mg cholesterol’ in the 1990s, 123 in the 2000s, and 0 before that).

In these cases, no proper semantic shifts occurred: the word ‘*indirectly*’ still had the same general meaning in the 1990s, and the word ‘*mg*’ in the 1990s and 2000s. However, the PRT/APD method indeed detected anomalous context variances in the time bins under analysis. Another interesting case belonging to this type is the word ‘*neutral*’ also appearing in Table 6.12: it seems that its 2000s burst is caused by the emergence of the frequent collocation ‘*gender* *neutral*', which is missing (or extremely rare) in the previous decades. Are we observing a new sense gradually appearing, or is it just contextual fluctuation? Anyway, independent of whether these variances are due to real changes in the word usage at these decades (caused by social and cultural developments) or due to improper corpus collection procedure, they are still objective bursts in the data. In this respect, this type of controversial predicted changes is different from those like '*progressive*' or '*designate*'. To some extent, this is another manifestation of a larger NLP problem of domain sensitivity (Okurowski, 1993): essentially, what PRT/APD detected was a domain change in comparison to overall genre structure of COHA.

Finally, the word '*Banish*' belongs to the proper names subset of the same 2nd class. It features clearly separated cluster of token embeddings containing exclusively the 1990s occurrences (bottom of Figure 6.14). In fact, all of them are mentions of 'Banish' as the name of one of the main characters of the 1996 novel 'The Standoff' by Chuck Hogan, for example:

(10)

1. '**Banish** slipped deeper into thought.'
2. '**Banish** smiled weakly at the sentiment.'
3. 'The sound man eyed him as he stepped inside, saying nothing about **Banish's** burnt face.'

The novel is included in COHA almost in its entirety, obviously bringing in a lot of '*banish*' usages very different from its mainstream verbal meaning.[23] This leads to the high change score we observe when comparing 1990s against 1980s: 0.7940, a strong burst compared to 0.7329 (1960s-1970s) and 0.7305 (1970s-1980s). Note that the change score is high again when looking at 2000s versus 1990s (0.7928). The obvious reason is that the 2000s corpus does not mention Banish from 'The Standoff' at all, so the meaning of '*banish*' has returned to its pre-1990s state (more or less equally distributed between the sense of 'TO EXPEL' and the sense of 'TO DESTROY, TO END').

Using '*Banish*' in this way is certainly creative, and even more importantly, these occurrences obviously denote something different from the regular meaning of '*banish*'. Of course, it can be disputed whether using a verb (or a common noun) as a proper name *is* coining a new sense. Note, however, that a very similar case of the word '*apple*' acquiring the new sense of a well-known company proper name is often used as a classic example for word sense disambiguation Manion (2014). From this point of view, '*banish*' certainly temporarily acquired a new sense based on the COHA 1990s corpus (without losing its old mainstream sense), thus constituting a proper diachronic semantic shift.

The take-away message here is that, when measuring the strength of semantic change with contextualized embeddings, one should watch out for the unexpected

[23] We use lemmatized and lower-cased corpora. In this case, pre-processing decisions can help: keeping proper names capitalized will avoid them mixing with common words.

results described above. The 1st class (words with naturally 'fluid' meaning) is clearly erroneous and ways must be devised to filter out these cases. Possible approaches to do this could include measuring change scores between random subsets of one and the same time bin: if they are as high as those between different time bins, the possible reason is the word's fluidity.

The 2nd class can be considered erroneous or not, depending on one's definition of semantic change (e.g., whether it includes context variance). It can be looked at as a training corpus problem: COHA is not entirely well-balanced with respect to sense distribution. On the other hand, any dataset is biased and incomplete, and the notion of a '100% balanced' corpus is in fact ill-defined without further refinement (balanced *for what*?). Arguably, the creators of COHA did not set an aim to somehow 'properly represent' the distribution of word senses (even if there existed robust methods to implement this).

But this also raises complicated questions about the nature of meaning and of what exactly it is to undergo a 'meaning shift', especially when we observe a case of context variance. If we stick to the distributional view that 'senses are in fact clusters of corpus usages' (Kilgarriff, 1997), these cases should definitely count as sense inventory changes, or at least the appearance of short-term senses which then fade away. Then, if one does not employ some external data sources (like ontologies or diachronic dictionaries), there is no reliable way to discern 'meaning changes' from 'differences in the underlying textual data': they are simply the same thing. This is an inevitable consequence of accepting the data-driven distributional paradigm, something we already noted in Chapter 2, when describing the shortcomings of this approach to semantics.

During our manual analysis, we also observed multiple cases where token embedding clusters of an unambiguous word manifested this word being used in different syntactic roles: the 3rd class of controversial change predictions. The example for the word '*phone*' is shown in Figure 6.15. There are three clusters of ELMo token embeddings, stable across all four decades. It turns out they group occurrences not on *semantic*, but more on *syntactic* grounds:

1. The top cluster contains sentences where '*phone*' is used as a subject:
   - '*Then the **phone** rang.*'
   - '*The **phone** yanked me awake.*'
2. Bottom left cluster contains sentences where '*phone*' is used as an object or as an oblique argument:
   - '*Hannah took a deep breath and grabbed the **phone**.*'
   - '*While you're away, talk to your son on the **phone**.*'
3. Bottom right cluster contains sentences where '*phone*' is used as a modifier part of compound nouns:
   - '*Please include a daytime **phone** number.*'
   - '*The **phone** calls from all over the the US have been so frequent.*'

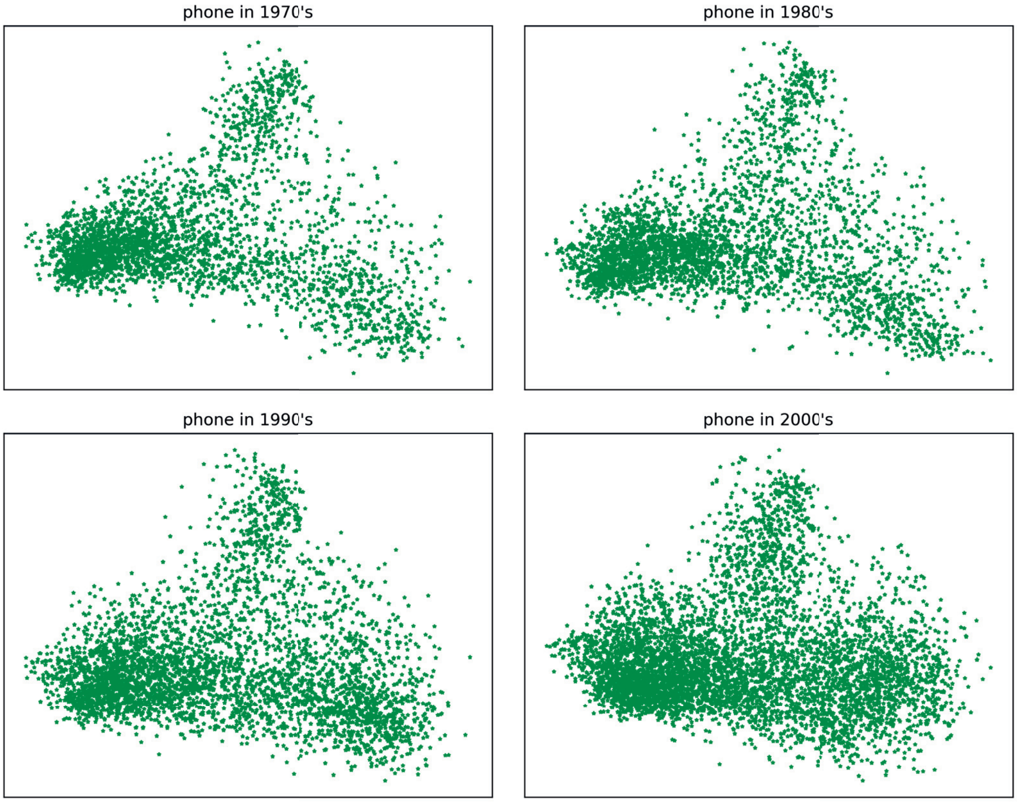


Figure 6.15: PCA projections of ELMo representations of each occurrence of the word '*phone*' in four different decades: stable syntactic clusters.

One can imagine that if for some reason the syntactic role distribution of a particular word changes diachronically, the semantic change detection methods based on contextualized embeddings would be triggered by this. As a result, a syntactic shift will be taken for a semantic one. '*Traditionally*' from Table 6.12 is such an example: for some reason, the 1990s COHA sub-corpus contains much less usages of this word as an adjective modifier ('*traditionally christian*', '*traditionally male*', etc) than the other decades, but there are no semantic changes. Interestingly, this syntactic influence on the resulting embeddings is expressed even though we extracted representations from the *top layer* of the neural network, which was shown by Peters, Neumann, Zettlemoyer, et al. (2018) to mostly contain *semantic* information.

We formulate the take-away message here as follows. Although contextualized embeddings like ELMo are indeed promising for the tracing of diachronic semantic shifts (especially for finding supporting examples from the corpus), their usage is not entirely straightforward. Contextualized representations are by definition very much influenced by contexts (especially 'exotic' ones) and fluctuations in corpus balance. They also often merge together syntactic and semantic characteristics of words. This can lead to a situation when a word occurrence receives a very different embedding not because the word has acquired a new sense, but because it is used in an unusual syntactic role, or because it is surrounded by unusual neighbors (for example, when the domain of the underlying texts has changed). Since the resulting semantic change score is a derivative of the arrays of token embeddings, one observes strong bursts which manifest changes in context variance of a word, not a semantic shift in the lexicographic meaning of this term. This is probably not what a historical linguist expects to see, although it can depend on the particular study and the working definition of 'semantic shift'.

Words with context-dependent 'fluid' meaning (like '*progressive*' or '*designate*' above) are another problem, as they will always exhibit strong change without it being of any significant linguistic interest. Finally, contextualized embeddings often merge together syntactic and semantic characteristics of words, which can be problematic as well. The discussed issues do not depend on a particular training algorithm, and there is no reason for it to not manifest itself also when using BERT and any other contextualized architectures (although to properly test it empirically could be an interesting future work).

It is not immediately clear whether improving the quality and representativeness of diachronic corpora can help alleviating this (producing more historical data is often not feasible or even impossible). In some cases, very simple preprocessing decisions can help: for example, keeping proper names capitalized in the corpora will address the issue of them mixing with common nouns. Another hypothetical remedy for some of the mentioned issues is smart handling of syntactic information from the representations used for calculating semantic change scores. This might be achieved by learning an optimal weighted function of different layers of the model in the process of training a binary classifier (shift or not shift) on manually annotated data.

6.6 Summary

This chapter described our experiments with employing recently introduced deep contextualized embedding architectures for lexical semantic change estimation. Our results for the SemEval-2020 Shared Task 1 (Subtask 2) and for the GEMS test set show that using contextualized embeddings to rank words by the degree of their semantic change produces strong correlation with human judgments, outperforming static embeddings. Models pre-trained on large external corpora and fine-tuned on the historical test corpora produce the highest correlation results, with ELMo (Peters, Neumann, Iyyer, et al., 2018) slightly but consistently outperforming BERT (Devlin et al., 2019) as a contextualizer. Considering that ELMo models have about half the number of parameters compared to BERT, we believe our results give a chance for NLP practitioners to make a more informed decision about which architecture to use in their cases.

Inverted cosine similarity between averaged contextualized embeddings and the average pairwise cosine distance between contextualized embeddings turned out to be the best semantic change detection methods. An interesting finding is that the former method favors the test sets with uniform gold score distribution, while the latter works best with the test sets where the gold score distribution is skewed towards low values. This distinction is not related to the language of the test set. We believe this dependency between the statistical properties of gold scores and the performance of semantic change detection systems deserves to be investigated further in future work. For the time being, we found that in a realistic case of not knowing the gold score distribution beforehand, one can use the average of these two methods' predictions (model ensemble), which proved to be a robust choice across the board, with the highest average correlation. Qualitatively, the proposed method confirm known semantic shifts (for example, the word '*cell*' acquiring the 'MOBILE PHONE' sense in the 2000s).

Additionally, we showed that the diversity of ELMo contextualized token embeddings for a particular English word in a given corpus does correlate with the number of the WordNet synsets for this word, and thus with the degree of its semantic ambiguity. Using this measure, we undertook an exploratory large-scale analysis of semantic change across five decades of the 20^{th} and 21^{st} centuries, and across three frequency tiers. We sorted the decades from the 1970s to the 2000s by their influence on the overall lexical ambiguity, and showed that the general tendency for the representation diversity to increase holds across all of these time bins, but most of it is manifested only when using incrementally trained contextualized embeddings. For this reason, we came to the conclusion that using incremental contextualized embeddings is generally not recommended, and single models should be used whenever possible, since they allow one to avoid the extra training bias.

At the same time, important issues were discovered and described. They are related to the fact that contextualized architectures capture many different word aspects, along with is semantics. This leads to token embedding changes captured by our methods but not representing proper semantic shifts (acquiring or losing a lexicographic sense). They are caused by the word occurrences being used in unusual environments and by imbalanced test corpus data (thus, increasing the size of the corpora will arguably not help). We identified three typical cases when high semantic change score is produced by our method, but it does not look like a 'proper' semantic shift:

1. Words of very context-dependent 'fluid' meaning, used in all sorts of contexts.
2. Words frequently used in a very specific context in a particular time bin ('data bursts'). There is a notable subclass here:
   - words used as a proper name in a particular time bin only.
3. Words being used in significantly different syntactic roles.

Contextualized architectures are unfortunately more vulnerable to these issues than the static architectures: precisely because in them, final word representations depend on the input context. This is not necessarily a problem for the case 2: new senses are born by using old words in new contexts, and clusters of irregular usages can be looked at as emerging senses. From this distributional perspective on 'senses', contextualized embeddings produce very reasonable representations. At the same time, the words from the cases 1 and 3 are erroneous semantic change 'hits', and we proposed some ways to remedy this. Researchers must take the aforementioned lexical groups into account when working with contextualized architectures for semantic change detection. In the future, ways should be devised for distinguishing word tokens used in different lexicographic senses from word tokens used in contextually varied surroundings, but in the same sense. Together with the 'sensitivity' of contextualized embeddings and semantic change detection methods to various aspects of word usage, this can potentially help to model the nature of subtle semantic shifts (of course, if they are manifested distributionally at all).

Another interesting issue is whether the proposed methods can be used for *classification* (actually detecting shifts). In this chapter, we dealt with the *ranking* task (SemEval-2020 Shared Task 1 Subtask 2), where one has to rank the word by the degree of their semantic change. Subtask 1, instead, challenges a system to tell whether a word has experienced a sense change (acquiring a new sense or losing an old one) or was stable across the time bins under analysis. This is a binary classification task, and we did not try to employ contextualized embedding based methods to solve it. However, possible approaches can be easily seen: from simply finding an optimal threshold of semantic change score (after this threshold is exceed, the word is considered to be shifted) to using various flavors of clustering token embeddings into groups corresponding to senses, and then detecting the emergence of new clusters. This can be potentially problematic: we have shown in Section 6.3 that the number of token embedding clusters produced by widely used algorithms like Affinity Propagation does not correlate well with the number of senses the word actually has. This can be addressed, for example, by using graph-based clustering, after converting vector representations into ego-graphs of the nearest neighbors (Logacheva et al., 2020).

Anyway, empirical results give us ground to state that despite all the challenges, approaches based on contextualized distributional embeddings are bound to replace traditional 'static' embeddings in diachronic semantic change modeling, as has already happened in several other natural language processing areas. Meaning in human languages is contextual and any attempts to build context-independent representations will always lead to severe over-simplifications. Higher expressiveness of contextualized models will allow researchers to come up with more persuasive examples and to develop change detection methods which will determine the *nature* of semantic shifts (narrowing, widening, metaphorization, metonymization etc). This will certainly remain an important direction of our own future work.

Chapter 7

Conclusion

Distributional semantic representations (word embeddings) trained on large amounts of linguistic data capture many aspects of word meaning. Due to the widespread use of word embeddings in natural language processing, a better understanding of these representations is of vital importance. There is a vast amount of literature on testing their abilities in a *synchronic* setup. However, at the time of commencing our work on this thesis,there was a definite lack of corresponding studies focusing on *diachronic* processes. If word embeddings are able to infer word meaning at a given point in time, they provide a good starting point for research aimed at modeling changes which this meaning undergoes over time. Such representations form a strong empirical basis for linguistic hypotheses testing and may give answers to many questions regarding lexical semantic change.

In this thesis, we analyzed the usage of distributional semantic representations for modeling of various types of diachronic semantic change. Diachronic word embeddings have by now become established as a central tool in the field of unsupervised semantic change detection. However, these architectures capture different aspects of natural language semantics, and semantic change manifests its multiple aspects in different linguistic phenomena. Semantic shifts proper are cases of one and the same word form acquiring a new lexicographic sense (like the English '*cell*' and the sense of 'MOBILE PHONE') or losing an old sense (like the German '*Zufall*' and the sense of 'SEIZURE') over time. Drift in context variance (like with the English 'distancing' at the time of COVID-19 pandemic) is another type of diachronic semantic change. Encyclopedic meaning or 'world knowledge' associated with a word can also change over time, without introducing new senses: this often happens with country or political group names, sometimes drastically changing their connotational semantics (cf. the word '*Crimea*' after the Russian annexation of the peninsula in 2014). Finally, diachronic semantic processes differ in the longitudes of the time spans they occur in.

The role of this thesis is to create a comprehensive (although not exhaustive) publication which covers a broad range of different types of semantic change captured by word embeddings, but at the same time employs consistent terminology and vision. We also made an attempt to provide a coherent story unfolding from simple to more complicated issues and from foundations to practical approaches. The thesis is summarizing the large body of research carried out by the author throughout the doctorate years. When the work on the thesis started in 2015, the field of diachronic semantic change detection was much more sparse than it is today. In particular, the use of dense word embeddings for this task was far from common, with only a few quite disconnected papers exploring this vein of research. Since then, the field grew much more mature.

This thesis has developed along with the field, as our research was conducted and published in peer-reviewed venues (14 main conference papers and journal articles, and nine workshop papers on the topic of semantic change and/or the properties of distributional meaning representations).

Overall, we design and test various methods to *probe* semantic word representations for diachronic information. Word embeddings are not directly interpretable by humans, making them to some extent a 'blackbox' (Linzen et al., 2019). One can probe embeddings for many different aspects of linguistic knowledge. We here explore their ability to capture *language change*: more specifically, temporal changes in various aspects of lexical semantics, including semantic relations between words. Note that although we cover other approaches to the semantic change detection task, word embedding-based methods still remain our primary focus throughout the thesis.

We first surveyed and systematized a large body of related work on the topic, both prior and concurrent to our work on this thesis. A selection of these methods (including those proposed by us) was used to investigate the semantic change of a linguistically defined category of words, namely evaluative adjectives, over time. We also introduced novel ways to evaluate semantic change detection methods and probe diachronic word embeddings: namely, through distant supervision from historical armed conflict datasets created by social scientists. The use of these datasets allows us to overcome the lack of gold standard semantic change data. Based on this foundation, we traced how such real-world event dynamics (in this case, armed conflict events) are captured by temporally-aware word embedding representations and how they are related to diachronic semantic change in named entities (for example, country names). In addition, we proposed novel methods (based on learning and re-applying optimal linear transformations) which use such representations to trace temporal dynamics of semantic relations between words.

Finally, we evaluated the potential of contextualized word embedding models like BERT and ELMo in tracing semantic change. In particular, we conducted extensive experiments with the methods employing such architectures on manually annotated semantic shift datasets: both well-established like GEMS (Gulordava and Baroni, 2011) and more recent like SemEval-2020 Shared Task 1 (Schlechtweg, McGillivray, et al., 2020). We showed that using contextualized architectures can significantly improve the performance of unsupervised semantic change detection in comparison to using static word embeddings. We also rigorously analyzed some unexpected and potentially controversial predictions of such methods. A linguistically motivated categorization of these issues was proposed with suggestions on how researchers can handle them.

Below, we summarize our main contributions and provide a more detailed summary of the results of our work. We will also outline directions of possible future research.

7.1 General contributions

Briefly, the main contributions of this thesis can be listed as follows:

1. We systematically show how word embedding-based methods are effective in diachronic semantic change detection. The thesis has proposed such methods for approaching several different aspects of semantic change modeling: ranging from datasets, tasks and evaluation measures to architectures and embedding training strategies.

2. We demonstrate how these methods tackle both semantic shifts proper and more subtle changes in context variance (still belonging to the domain of semantics).

3. We have structured unsupervised methods for semantic change modeling along several axes to allow for meaningful comparison between different approaches. The most important findings and events in the field are outlined and discussed.

4. In one case study, we apply some of the established semantic change detection methods to investigate the dynamics of semantic change that evaluative adjectives undergo over time. This is explored for three languages.

5. We demonstrate how non-linguistic temporally annotated datasets (in this case, containing armed conflicts data) can be used to probe diachronic word embeddings.

6. We introduce a novel extension of diachronic analogical reasoning, and propose and evaluate a model for approaching this task.

7. The previous point indicates that diachronic word embeddings capture information about temporal changes in word relations, not only single words.

8. We propose and evaluate several new methods for semantic change detection based on contextualized embedding architectures. These methods outperform previous static embedding-based approaches on several test sets and languages.

9. We find that contextualized methods allow easier inspection and visualization of temporal shifts in word meaning. At the same time, they are prone to producing controversial predictions (a high semantic change score is produced, but it is not a real semantic shift in the lexicographic sense) in some cases. These cases are identified and manually categorized.

7.2 Summary

Our primary research question (*RQ0*) was formulated as follows: **is it possible to reliably model diachronic semantic change using dense distributional word representations?** We addressed *RQ0* through several case studies across several languages. These case studies involved working with time-annotated corpus data, preparing gold standard datasets, training and comparing different flavors of diachronic word embeddings, and evaluating semantic change detection methods (both quantitatively and qualitatively).

As a result, we showed that the answer to *RQ0* is partially positive: word embedding changes do capture diachronic semantic change. It is to some extent possible to create embedding-based computational systems that can model and detect various aspects of semantic change: shifts in the composition of words' lexicographic senses, slight but consistent drift in context variance (often caused by short-term event dynamics), and even changes in typed semantic relations between words. One of the advantages of these methods is that it is possible to employ them in an unsupervised manner, with no labeled data. Even under such setups, methods based on dense distributional representations still yield meaningful results.

At the same time, many challenges still remain. Some of them are technical, like the issue of making diachronic embedding spaces comparable or the issue of controlling for word frequencies. Some of these challenges are conceptual, like the issue of different lexical aspects being expressed in a word's usage and then making their way into a word embedding together, without a clear way to differentiate, say, semantics from syntax.

Other research directions stem from *RQ0*. In the subsections below, we describe the outcomes of the thesis, following the order of the respective chapters and the research questions identifiers from the Introduction (Section 1.2).

7.2.1 State of the field

We started the thesis with surveying the current research related to unsupervised semantic change detection. This field is now developing very fast, and often in unexpected directions. In the recent two or three years, the field has undergone a strong increase in the number of papers published and events organized. As new papers on the topic appear almost weekly, any attempt to exhaustively summarize the relevant research would not be realistic. However, in the present thesis, we have surveyed the existing publications which we deem to be most important. We believe that the structure and conceptualization offered in Chapter 3 will hopefully still be relevant for many years to come. Our survey on diachronic semantic change modeling using word embeddings (Kutuzov, Øvrelid, et al., 2018) is now a widely cited reference in the field (Chapter 3 is partially based on this paper).

To address our research question *RQ1.1* ('What are the main axes along which one can structure the current research?'), we covered the linguistic nature of semantic shifts (both semasiological and onomasiological), the typical sources of diachronic data for training and testing, and the distributional approaches used to model them: from frequency-based methods to static and contextualized word embeddings. The distributional methods were structured according to several axes: the nature of diachronic data, evaluation metric, the type of word embedding algorithm, model alignment approach, etc. We also emphasized the difference between diachronic semantic change detection algorithms being used for 1) assessing *more linguistically oriented questions* (as in Chapter 4), and 2) addressing *practical NLP or text analysis tasks* (as in Chapters 5 and 6).

As a way of addressing *RQ1.2* (‘What were the primary discoveries in recent years?’), we outlined the most important findings and events in the field of unsupervised semantic change detection up to 2020 as a timeline in Figure 3.6. It is important to note here that while the survey in Chapter 3 is up-to-date as of 2020, other parts of the thesis were carried out step by step in different time periods since the beginning of this PhD study in 2015. Most of the thesis chapters are rooted in our papers published in 2017–2020. These studies were significantly reworked and expanded before inclusion into the thesis.

The emerging field of semantic change detection is still relatively new, and although recent years has seen a string of significant discoveries and academic interchange, much of the research still appears slightly fragmented. This survey is partly aimed at addressing this issue and presenting computational detection of diachronic semantic shifts with word embeddings as a more coherent story.

7.2.2 Diachronic evolution of a linguistically defined category

Research question *RQ2.1* (‘Do evaluative adjectives change over time faster than other types of adjectives?’) is linguistic in its nature. It serves as an example of a linguistic case study involving methods of unsupervised semantic change detection we described before that. Particularly, in Chapter 4, we measured the intensity of diachronic semantic change in adjectives across three languages (English, Norwegian and Russian) and five decades (1960s, 1970s, 1980s, 1990s, 2000s), to test whether evaluative adjectives change faster than other adjectives. This research was motivated by several well-known examples of English adjectives becoming evaluative or changing their polarity within comparatively short periods of time (cf. ‘*terrific*’, ‘*sick*’, etc).

We did not propose any new models in this part of the thesis, but tested the applicability of some of the existing ones to a concrete and linguistically motivated problem limited to a well defined lexical category. Our results showed that, contrary to the initial intuition, evaluative adjectives change over time *slower* (statistically significant at $p < 0.1$), if we measure change as the mean of pairwise differences between successive decades, and not as a steady drift in one particular direction. At the same time, when measuring the probability of steadily ‘shifting’ from an original meaning across time, in our experiments evaluative adjectives *do not differ from other adjectives at all* (on any statistically significant level). Thus our answer to *RQ2.1* is negative. There is no statistical evidence for evaluative adjectives to undergo faster diachronic semantic change, at least with the observation window of five decades. This holds for three different languages:

English, Norwegian and Russian. This research was originally published as Rodina, Bakshandaeva, et al. (2019).

These observations are not frequency artifacts, since we observe the same behavior when controlling for word frequencies. The controlled experiments additionally allowed us to trace how word embedding-based semantic change detection methods are influenced by frequency in different ways. In particular, for frequent words, Jaccard distances between the nearest neighbors and cosine distances between Procrustes-aligned models tend to yield *lower* semantic change scores, while the Global Anchors method tend to yield *higher* change scores.

We also conducted an additional experiment with the increased 'observation window' of 10 decades for English (starting from the 1910s). In this case, we observed a more expressed steady shift in one particular direction for evaluative adjectives (but still less expressed for the pairwise differences between successive decades). Our interpretation is that there is no difference between evaluatives and other adjectives in their short-term fluctuations (independent of the width of the observation window, be it five decades or 10). However, if we observe language data for a longer time, diachronic embedding-based methods may start to capture and show consistent movement of evaluative adjectives away from their original meaning.

7.2.3 Evaluating semantic change detection through extra-linguistic data

After applying a selection of existing semantic change detection algorithms in a linguistic case study, we moved on to actually evaluating them. The goal of the subsequent chapter was to probe what kind of information about cultural semantic change is captured by diachronic word embeddings.

The field of automated diachronic semantic change detection often suffers from the lack of manually annotated data, even now in 2020. This problem was even more expressed at the onset of the work on this thesis. We proposed to compensate this by using other temporally annotated datasets (not necessarily of purely linguistic nature) and assuming that the real-world historical events described in these datasets are strongly correlated with diachronic changes in word meaning. In particular, we used the Uppsala Conflict Data Program (UCDP) datasets which includes start and end dates of armed conflicts throughout the world.

To answer research question *RQ3.1* ('Can external datasets be used as proxies to evaluate change detection methods?'), a version of the UCDP Conflict Termination dataset was created, linguistically pre-processed and published (we call it Armed Conflict Evaluation Test Set). Using this historical armed conflict dataset, we showed that it is indeed possible to predict real-world events based on word vector changes in models trained on news texts. Moreover, it is possible to evaluate the semantic shift detection algorithms themselves, based on how good they are in reflecting the dynamics of real-world data. This data can be used as a source of extra-linguistic indicators useful for evaluating semantic change detection methods. This represents a case of distant supervision – a strategy that is becoming increasingly used in natural language processing (Fang and Cohn, 2016).

By evaluating these methods for their ability to detect or predict changes in the real world, we were able to better understand what information about temporal changes in connotational word meaning is captured by distributional word embeddings (thus, probing them). Several different approaches to extract this information were tested and evaluated in Chapter 5 using the Armed Conflict Evaluation Test Set.

Note that semantic change in this part of the thesis (unlike Chapter 4 and Chapter 6) is mostly concerned with 'world knowledge', and occurs for proper names. This corresponds to the context variance span on the semantic proximity scale: the words dramatically change their typical contexts without changing their lexicographic senses. We claim that such changes are still semantic, although they are different from what we have dubbed semantic shifts proper.

One can use semantic change detection methods in a comparatively simple setup when one measures the temporal drift of a geographical location embedding in relation to conflict domain specific 'anchor words' like '*kill*', '*casualty*', etc. This allows us to detect the start or end of an armed conflict based only on the analysis of word vector changes which in turn reflect changes in context variance of a particular named entity. For example, in 2006, the vector representation of '*Congo*' becomes much closer to these anchor words by cosine similarity (in comparison to 2005). The ultimate reason for this is the fact that armed conflicts resumed in this country this year. This event influenced the connotational components of the meaning of '*Congo*', because the 'world knowledge' associated with this word has changed. This phenomenon is captured by diachronic embedding-based semantic change detection methods. We described such experiments in Section 5.2, achieving reasonable performance on predicting armed conflicts from the Gigaword news texts, by comparing the similarity of country embeddings to manually or automatically selected anchor words related to war and peace. This approach significantly outperformed the frequency baseline. Parts of this work were previously published as Kutuzov, Velldal, et al. (2017b).

7.2.4 Diachronic dynamics of semantic relations

The majority of research on diachronic semantic change focuses on shifts in meaning which occur to single words (or other singular linguistic entities). Continuing to rely on armed conflict datasets as evaluation data, in Section 5.3, we investigated how diachronic word embeddings can serve as the foundation for systems which are able to trace the change of semantic *relations* over time. In comparison to single words, relations are more complex and high-level structures. Single words or other entities function as their parts. This problem, indicated in our research question *RQ3.2* ('Do word embeddings capture information about diachronic changes in semantic relations?'), is similar to the well-known word analogies task, but is more difficult and subtle than single-word semantic change modeling, since it involves the analysis of entity tuples (or even triplets or quadruplets). Similar (but not identical) tasks were previously called 'diachronic analogies' (Orlikowski et al., 2018) or 'temporal analogues/analogies' (Tahmasebi, Borin, and Jatowt, 2018).

In this work, we considered the task of detecting and predicting armed groups active in particular geographical locations. This is essentially answering questions like 'Does this semantic relation still hold between the entity X and the entity Y after some time has passed?', where X is, for example, '*India*', and Y is '*United Liberation Front of Assam (ULFA)*'. This setup fuses both *onomasiological* and *semasiological* changes. Parts of this research were previously published as Kutuzov, Velldal, et al. (2017a) and Kutuzov, Velldal, et al. (2019).

We addressed this task by learning linear transformations (projections) on incrementally trained diachronic word embeddings. In sections 5.3 and 5.4, we found that this approach significantly outperforms the baselines (Vector Offset and projections on Procrustes-aligned embeddings) and can even be applied in cases of one-to-zero and one-to-many relations. In particular, it successfully predicts the state of a semantic relation at a given time period, based purely on word embeddings trained on the news texts published in this time period and on the manually annotated data about similar relations in the past (even if the relation participants were completely different). We also showed how the comparatively simple technique of cosine thresholding can be used to significantly decrease the amount of false positive answers produced by this approach (when an insurgent group is predicted for a country with no armed conflicts at all).

In sum, we have introduced a novel extension of the task of analogical reasoning (adding a temporal dimension and open-ended relations) in addition to proposing a model for approaching this extended task. We found that geometric directions in diachronic word embedding models can correspond to very subtle semantic relations. Thus, we answer positively to *RQ3.2*, after demonstrating that these relations can be traced over time to detect whether they persist or die out, allowing us to conclude whether the relation still holds between the entities or not.

These experiments involved only one type of relation: that of an armed conflict. However, the approach of projection learning itself is relation-agnostic. It can potentially be used for any kinds of entities linked by any kind of one-to-X semantic connections which undergo change over time. Analyzing diachronic changes in semantic relations on the basis of word embeddings leads to findings far beyond the usual 'king is to queen is as man is to woman' analogy example by Mikolov, Yih, et al. (2013). Arguably, relations of the kind we have studied here are more challenging because:

1. the entities can be in one-to-X relations to each other;
2. the entities' involvement in relations can depend on the time period;
3. the relations themselves can change their form and nature (for example, transforming from one-to-one to one-to-many).

Our experiments with the Armed Conflict Evaluation Test Set described above allow us to give a positive answer to *RQ3.1*. We successfully demonstrated how an external dataset can be used as a proxy for evaluating semantic change detection algorithms based on word embeddings. This can be of help for languages still lacking proper semantic shift test sets.

7.2.5 Contextualized embeddings in semantic shift detection

The recently introduced deep contextualized meaning representations transform a word embedding model from a simple vector lookup table to a full-fledged language model (based on recurrent neural networks, transformers, etc.). Linguistically, this means that different representations are inferred for different tokens of the same word type in different contexts.

In Chapter 6, we described our experiments with lexical semantic change estimation based on contextualized embedding architectures. Four methods employing such architectures were proposed. Addressing the research question *RQ3.3* ('What new perspectives do contextualized architectures bring to semantic change detection?'), we showed how these methods allow for easier inspection and visualization of temporal shifts in word meaning. The reason for this is that it is now possible to position all the occurrences of a word (token embeddings) in a vector space where occurrences with similar semantics are close to each other and vice versa. This brings a natural grouping or clustering of a word's occurrences according to its different word senses. As these groupings change over time, the dynamics of new senses being born and old senses going extinct can be observed. Additionally, this greatly simplifies the task of collecting corpus examples relevant to emerging or disappearing word senses.

To answer our research question *RQ3.4* ('Do contextualized embeddings outperform static embeddings in this task?'), we conducted an empirical evaluation of the proposed methods. Our results for the SemEval-2020 Shared Task 1 Subtask 2 (Schlechtweg, McGillivray, et al., 2020) and for the GEMS (Gulordava and Baroni, 2011) test sets showed that using contextualized embeddings to rank words by their degree of semantic change produces strong correlations with human judgments, outperforming previous methods based on static embeddings. Models pre-trained on large external corpora and fine-tuned on historical test corpora produce the highest correlations, with ELMo (Peters, Neumann, Iyyer, et al., 2018) slightly but consistently outperforming BERT (Devlin et al., 2019) as a contextualizer. Considering that ELMo models have about half as many parameters than BERT, we believe our results give a chance for NLP practitioners to make a more informed decision about which architecture to use in their cases. These results hold for English, German, Latin and Swedish.

We further introduced several new approaches for estimation of semantic change, based on contextualized embeddings. Inverted cosine similarity between averaged token embeddings (so called 'PRT') and the average pairwise cosine distance between token embeddings (so called 'APD') turned out to be the best semantic change detection methods, yielding the highest scores for the particular languages in SemEval-2020 Task 1 Subtask 2 test set. An interesting finding is that the former method favors test sets with uniform gold score distribution, while the latter works best with test sets where the gold score distribution is skewed towards low values. This distinction is not related to the language of the test set. We found that in a realistic case of not knowing the gold score distribution beforehand, one can use the average of these two methods' predictions (a model ensemble), which turned out to be a robust approach across the board, with the highest average correlation. At the time of writing, the results of this method outperform all publicly available SemEval-2020 Shared Task 1 submissions. In a significantly shortened version, these experiments were described in Kutuzov and Giulianelli (2020).

Additionally, we showed that the diversity of ELMo contextualized token embeddings for a particular English word in a given corpus does correlate with the number of the WordNet synsets (senses) for this word, and thus with the degree of its semantic ambiguity. Using this measure, we undertook an exploratory large-scale analysis of semantic change across five decades of the 20^{th} and 21^{st} centuries, and across three frequency tiers. As a result, we came to the conclusion that using incrementally trained contextualized embeddings is generally not recommended, and single models trained from scratch on the full available corpus should be used whenever possible. They allow us to avoid the extra training bias which is manifested in a model yielding more and more diverse token embeddings as it is trained on more and more data.

Another important issue related to *RQ3.3* is the fact that contextualized architectures capture not only token semantics, but also syntactic (and maybe pragmatic) features of word tokens. This leads to token embedding changes captured by our methods but not representing semantic shifts proper (acquiring or losing a lexicographic sense). These phenomena are caused by the word occurrences being used in unusual environments or by imbalanced test corpus data. We identified three typical cases when a high semantic change score is produced, but where it does not correspond to a real semantic shift in the lexicographic sense:

1. Words of very context-dependent 'fluid' meaning, used in all sorts of contexts.

2. Words frequently used in a very specific context in a particular time bin ('data bursts'). There is a notable subclass here:

    - words used as a proper name in a particular time bin only.

3. Words being used in a significantly different syntactic role in a particular time bin.

Contextualized architectures are unfortunately more vulnerable to these issues than the static architectures: precisely because in them, final word representations depend on the input context. This is not necessarily a problem for case 2: new senses are born by using existing words in new contexts, and clusters of irregular usages can be looked at as emerging senses. From this distributional perspective on 'senses', contextualized embeddings produce very reasonable representations. At the same time, the words from cases 1 and 3 are erroneous semantic change 'hits', and in this thesis, we proposed some ways to remedy this situation.

Still, empirical results support the point of view that the approaches based on contextualized distributional embeddings are bound to replace traditional 'static' embeddings in diachronic semantic change modeling, as has already happened in several other natural language processing areas. Meaning is contextual, as was noted by linguists a long time ago (Firth, 1935); any attempts to build context-independent representations will always lead to severe over-simplifications. The higher expressiveness of contextualized models will allow researchers to come up with more persuasive examples and to develop change detection methods which will hopefully further determine the *nature* of semantic shifts (narrowing, widening, metaphorization, metonymization, etc).

In sum, the answer to *RQ3.4* is that contextualized embeddings do outperform static embeddings in lexical semantic change detection. Considering the many technological advantages of contextualized architectures, they are certainly worth trying for any practitioner interested in the task of semantic change modeling.

7.3 Future work

Unsupervised semantic change detection (including using word embeddings) is in general far from being a solved problem. The field is still young and has a considerable number of open challenges. We consider most of these challenges as part of our future work. We briefly describe some of them below.

- The existing methods should be expanded to a *wider scope of languages*. Hamilton, Leskovec, et al. (2016b) and other authors have started to analyze other languages, but the overwhelming majority of publications still apply only to English corpora. This thesis also addresses the issue to some extent in Chapter 4, providing comparative analysis for three languages and in Chapter 6 evaluating our methods on four languages. The language coverage should be expanded to include more typologically and genetically diverse languages and more varied time spans.

- Carefully designed and robust *gold standard test sets* of semantic shifts (of different kinds) should be created and made publicly available. This is a difficult task in itself, but the experience from synchronic word embeddings evaluation (Hill et al., 2015) and other NLP areas hints that it is possible. The SemEval-2020 Task 1 dataset (Schlechtweg, McGillivray, et al., 2020) based on the DuREL framework (Schlechtweg, Schulte im Walde, et al., 2018) is a great example. Additional such datasets should be created and compared, with more languages and language families covered. It is also important to study the differences between the existing datasets and properly analyze the issues like the one we noted in Chapter 6: the dependency between the statistical properties of gold scores in the test sets and the performance of semantic change detection systems deserves to be investigated further.

- Historical corpora are the ultimate source of data for the field. The existing results can be refined and improved using larger or cleaner (with pre-processing artifacts fixed) text collections, for example, the recently presented Clean COHA (Alatrash et al., 2020).
- There is a need for more explanatory *formal mathematical models of diachronic vector representations of meaning*. They should make clear what deficiencies of the current representations need to be first addressed. Arguably, this will follow the vein of research in joint learning across several time spans, started by Bamler and Mandt (2017), Yao et al. (2018) and Rosenfeld and Erk (2018), but other directions are also open.
- Most current studies stop after stating the simple fact that a semantic shift occurs. However, more detailed analysis of the nature of the shift is needed. This includes:

    1. *Sub-classification of types of semantic shifts* (broadening, narrowing, metaphorization, etc). This problem was to some degree addressed by Mitra et al. (2014) and Tahmasebi and Risse (2017a), but much more work is required.
    2. *Identifying the source of a shift* (for example, linguistic or extra-linguistic causes). This causation detection is closely linked to the division between linguistic drifts and cultural shifts, as explained in Hamilton, Leskovec, et al. (2016a). Again, manually annotated datasets of both linguistically motivated and culturally motivated semantic shifts are needed.
    3. *Quantifying the weight of senses* acquired over time. Many words are polysemous, and the relative importance of senses is flexible (Frermann and Lapata, 2016). To address this, methods from sense embeddings research (Bartunov et al., 2016) might be employed. Another solution which is gaining popularity now is using contextualized embedding architectures, as shown in Chapter 6 of this thesis. This work is still very young, and much more is to be done here.
    4. *Identifying groups of words that change their meaning together* in correlated ways. Some research in this direction was started in Dubossarsky, Weinshall, et al. (2016), who showed that verbs change more than nouns, and nouns change more than adjectives. Also, in this thesis, we rejected the hypothesis that evaluative adjectives shift faster than other adjectives (Chapter 4). This is also naturally related to demonstrating the (non-)existence of the ‘laws of semantic change’ (see Section 3.3) and to studying the processes of co-lexification. Since the number of lexical groups is potentially infinite, most important and interesting groups should be identified and analyzed.

5. *Avoiding catastrophic forgetting* when training embeddings incrementally. Relational structures tend to completely change after significant updates of the model. Can it be avoided somehow, while still using new training data to the full extent? It is also possible that incremental training should be avoided: especially with contextualized embeddings, where there exists the alternative to use a single pre-trained model to infer token representations from time-specific corpora.

- The experiments with diachronic relation prediction in Chapter 5 involved only one type of semantic relations: that is, armed conflicts. However, our proposed approach of prototypical projection learning itself is relation-agnostic. It can be potentially used for any kinds of entities linked by any kind of one-to-X edges. We already know it can be employed in diachronic tasks. Provided we possess the relevant corpora, this potentially paves the way to automatically inferring the temporal dynamics of semantic relations between persons and organizations, ideas and technologies, etc.
- As we showed in Chapter 6, contextualized embedding-based methods demonstrate promising results in semantic change detection. However, ways should be devised for distinguishing word tokens used in different lexicographic senses from word tokens used in contextually varied surroundings, but in the same sense. We also intend to explore the possibilities to improve our best-performing methods (PRT and APD), in particular with respect to removing the outlier embeddings before calculating the semantic change score.
- This thesis was mostly about estimating the degree of semantic change and ranking linguistic entities according to this degree. However, we are also very interested in the task of binary classification of words into those that changed their meaning and those that did not. This is unavoidably linked with the issues of senses and their emergence or dying out.

Finally, community building is essential for any relatively young research field. As more publications related to the problem of automatic semantic change modeling appear, more venues will be needed for open discussion of the issues arising in this research area. The 1st International Workshop on Computational Approaches to Historical Language Change at the ACL2019 conference (Tahmasebi, Borin, Jatowt, and Y. Xu, 2019) and the SemEval2020 task for Unsupervised Lexical Semantic Change Detection[1] (Schlechtweg, McGillivray, et al., 2020) are perfectly timed steps in the right direction.

7.4 Publicly available code, datasets and models

All the code, datasets and trained embeddings produced in the course of the work on this thesis, are made publicly available online under permissive licenses.

[1] https://competitions.codalab.org/competitions/20948

Below we briefly enumerate the published items.

7.4.1 Code

- Python code for learning and evaluating linear projections between entities in diachronic word embeddings. The resulting prototypical relations are tested on other time periods.[2]
- Python code to measure the pace of semantic change of any given word in any given sequence of time spans using word embeddings and different semantic change estimation algorithms.[3]
- Minimal Python code to work with the vectors from pre-trained ELMo models in up-to-date TensorFlow versions.[4]
- Implementations of lexical semantic change detection algorithms using contextualized embeddings.[5]

7.4.2 Datasets

- Historical Armed Conflict Evaluation Test Set.[6]
- English, Norwegian and Russian lists of evaluative adjectives. We adapted these lists from external sources, by translating, reformatting, filtering and pre-processing them.[7]

7.4.3 Pre-trained models

- Diachronic word embeddings trained on the English COHA corpus.[8]
- Diachronic word embeddings trained on the English Gigaword corpus.[9]
- Diachronic word embeddings trained on the English News on Web (NOW) corpus.[10]
- Diachronic word embeddings trained on the Norwegian NBdigital corpus.[11]
- Diachronic word embeddings trained on the Russian National corpus.[12]

[2] https://github.com/ltgoslo/diachronic_armed_conflicts
[3] https://github.com/ltgoslo/diachronic_multiling_adjectives
[4] https://github.com/ltgoslo/simple_elmo
[5] https://github.com/akutuzov/semeval2020
[6] https://github.com/ltgoslo/diachronic_armed_conflicts/tree/master/2019_dataset
[7] https://github.com/ltgoslo/diachronic_multiling_adjectives/tree/master/datasets
[8] http://vectors.nlpl.eu/repository/11/188.zip
[9] http://vectors.nlpl.eu/repository/11/191.zip
[10] http://vectors.nlpl.eu/repository/11/192.zip
[11] http://vectors.nlpl.eu/repository/11/189.zip
[12] http://vectors.nlpl.eu/repository/11/190.zip

Appendices

Appendix A

Structure of the GED dataset

The structure of the GED dataset is rather rich. It contains information about:

- *event identifiers*,
- *actors and dyads* (established pairs of actors),
- *sources*,
- *geography*,
- *time*,
- *clarity* (whether the reporting was sufficiently clear for the coder to be able to fully identify the event itself or not),
- *fatality figures.*

For the full explanation, see the relevant Codebook (Sundberg and Melander, 2013).[1]

To illustrate the richness of the data, below we give an example of a GED dataset entry with comments where necessary. The entry describes the event of heavy shelling which took place in the Northern Sri Lanka in the March of 2009. This event is linked to the long-standing conflict between the government of Sri Lanka and the organization of Tamil Eelam. This structured information was extracted by the UCDP human coders from the news text 11:

(11) Embassy Colombo reported that 72 people were killed and 91 injured by continued shelling in the NFZ. The U.S. Embassy was told that a multi-barrel rocket launcher sent 40 shells into the NFZ in one barrage, and that 21 of the 72 deaths were individuals who were in line to receive their food ration. Upon learning of the shelling, an organization spoke with the GSL military in Vavuniya and requested that the shelling cease. An organization provided messages from a source in Mullaittivu with similar details about a multi-barrel rocket launcher attack in Mullivaikkal, wounding 93 people. A source near Mattalan reported to HRW very heavy shelling to the west. Many shells landed within 200 meters of the source.

1. **Event identifiers**

    a) *id*: '76479' (unique for each event)

[1] https://ucdp.uu.se/downloads/ged/ged191.pdf

   b) *year*: ‘2009’
   c) *active_year*: ‘True’ (the event belongs to an active conflict/dyad/actor year)
   d) *type_of_violence*: ‘1’ (1 means state-based conflict, 2 is non-state conflict, and 3 is one-sided violence)
   e) *conflict_new_id*: ‘352’ (unique code for the conflict)
   f) *conflict_name*: ‘Sri Lanka (Ceylon):Eelam’

2. **Actors and dyads**

   a) *dyad_new_id*: ‘776’ (unique identifier for a dyad: a pair of conflict sides)
   b) *dyad_name*: ‘Government of Sri Lanka - LTTE’ (the name of the dyad)
   c) *side_a*: ‘Government of Sri Lanka’ (state actors are as a rule assigned side *a*)
   d) *side_b*: ‘LTTE’ (‘Liberation Tigers of Tamil Eelam’ armed insurgent group)
   e) *side_a_new_id*: ‘145’ (actor identifier)
   f) *side_b_new_id*: ‘320’ (actor identifier)

3. **Sources**

   a) *number_of_sources*: ‘-1’ (-1 means not applicable here; this field has real values only for very recent data)
   b) *source_article*: ‘Report to Congress on Incidents During the Recent Conflict in Sri Lanka. U.S. Department of State, 2009.’ (the title of the source)
   c) *source_office*: no information
   d) *source_date*: no information
   e) *source_headline*: no information
   f) *source_original*: ‘Embassy Colombo’ (organization providing information about the event)

4. **Geography**

   a) *where_prec*: ‘2’ (geographical precision; the higher is this value, the less sure is the coder about the location; 2 means the coder is certain the event occurred within 25 km from the coded coordinates)
   b) *where_coordinates*: ‘NFZ2’ (standardized and normalized name of the location)
   c) *adm_1*: ‘Northern’ (administrative division where the event took place)

d) *adm_2*: ‘Mullaittivu’ (even more fine-grained administrative division, can be a village or a small town)

e) *latitude*: ‘9.321175’

f) *longitude*: ‘80.771919’

g) *geom_wkt*: ‘POINT (80.771919 9.321175)’ (An Open Geospatial Consortium textual representation of the location)

h) *priogrid_gid*: ‘143082’ (PRIO Grid cell of the event (Tollefsen et al., 2012))

i) *country*: ‘Sri Lanka’

j) *country_id*: ‘780’

k) *region*: ‘Asia’

5. **Clarity**

a) *event_clarity*: ‘1’. It means the coder was able to fully identify the event, lower values mean more uncertainty)

6. **Time**

a) *date_prec*: ‘1’ (date precision; 1 means the coder knows the exact date of the event)

b) *date_start*: ‘2009-03-11’

c) *date_end*: ‘2009-03-11’

7. **Fatality figures**

a) *deaths_a*: ‘0’ (always 0 for one-sided violence events)

b) *deaths_b*: ‘0’ (always 0 for one-sided violence events)

c) *deaths_civilians*: ‘0’ (number of deaths of persons of civilians)

d) *deaths_unknown*: ‘0’ (number of deaths of persons of unknown status)

e) *best*: ‘0’ (the sum of *deaths_a*, *deaths_b*, *deaths_civilians* and *deaths_unknown*)

f) *high*: ‘72’ (the highest reliable estimate of fatalities)

g) *low*: ‘0’ (the lowest reliable estimate of fatalities)

Appendix B

Per-year accuracies for next-year armed group predictions

In this appendix, we provide full list of per-year accuracies for our best-performing methods for next-year insurgents predictions from Chapter 5.

Table B.1 gives the details of the performance of the incremental training approach with vocabulary expansion and single-year projection learning. Its average accuracy was given in Table 5.6.

Tables B.2 and B.3 report yearly F1 scores for the same task in the One-to-X setup for the Gigaword and NoW datasets correspondingly. The average F1 scores for these experiments were given in Table 5.10 (Section 5.4).

Year	**Accuracy @1**	**Accuracy @5**	**Accuracy @10**
1996	0.317	0.439	0.512
1997	0.333	0.524	0.643
1998	0.415	0.634	0.634
1999	0.325	0.525	0.550
2000	0.348	0.478	0.565
2001	0.404	0.617	0.638
2002	0.366	0.488	0.585
2003	0.395	0.632	0.684
2004	0.351	0.514	0.541
2005	0.355	0.581	0.645
2006	0.300	0.500	0.525
2007	0.351	0.649	0.649
2008	0.417	0.583	0.639
2009	0.400	0.629	0.714
2010	0.516	0.710	0.871
Average	0.373	0.567	0.626
Standard deviation	0.054	0.077	0.090

Table B.1: Yearly accuracies (including OOV words) of predicting next-year insurgents based on location vectors and projections learned from the previous year. Best approach from Section 5.3.

	F1 scores	
Year	**Projection baseline**	**Threshold**
1996	0.22	0.29
1997	0.28	0.40
1998	0.34	0.51
1999	0.29	0.37
2000	0.30	0.48
2001	0.35	0.46
2002	0.27	0.39
2003	0.27	0.52
2004	0.23	0.42
2005	0.21	0.30
2006	0.24	0.38
2007	0.26	0.33
2008	0.29	0.40
2009	0.33	0.45
2010	0.32	0.47
Average	0.28	0.41
Standard deviation	0.04	0.07

Table B.2: Yearly F1 scores of armed conflicts prediction in the One-to-X setup (Gigaword dataset).

	F1 scores	
Year	**Projection baseline**	**Threshold**
2011	0.25	0.46
2012	0.29	0.50
2013	0.36	0.32
2014	0.37	0.34
2015	0.37	0.42
2016	0.33	0.36
2017	0.41	0.44
Average	0.34	0.41
Standard deviation	0.05	0.07

Table B.3: Yearly F1 scores of armed conflicts prediction in the One-to-X setup (NoW dataset).

Bibliography

Aggelen, Astrid van, Antske Fokkens, Laura Hollink, and Jacco van Ossenbruggen (2019). “A larger-scale evaluation resource of terms and their shift direction for diachronic lexical semantics”. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press: Turku, Finland, pp. 44–54.

Aitchison, Jean (2001). *Language change: Progress or decay?* Cambridge university press.

Alatrash, Reem, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde (2020). “CCOHA: Clean Corpus of Historical American English”. English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association: Marseille, France, pp. 6958–6966.

Amrami, Asaf and Yoav Goldberg (2018). “Word Sense Induction with Neural biLM and Symmetric Patterns”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Brussels, Belgium, pp. 4860–4867.

Artetxe, Mikel, Sebastian Ruder, Dani Yogatama, Gorka Labaka, and Eneko Agirre (2020). “A Call for More Rigor in Unsupervised Cross-lingual Learning”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Online, pp. 7375–7388.

Auguste, Jeremy, Arnaud Rey, and Benoit Favre (2017). “Evaluation of word embeddings against cognitive processes: primed reaction times in lexical decision and naming tasks”. In: *Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics: Copenhagen, Denmark, pp. 21–26.

Azarbonyad, Hosein, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps (2017). “Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM ’17. ACM: Singapore, Singapore, pp. 1509–1518.

Bakarov, Amir, Roman Suvorov, and Ilya Sochenkov (2018). “The Limitations of Cross-language Word Embeddings Evaluation”. In: *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics: New Orleans, Louisiana, pp. 94–100.

Bamler, Robert and Stephan Mandt (2017). “Dynamic word embeddings”. In: *International Conference on Machine Learning*, pp. 380–389.

Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell (2020). *A Twitter Dataset of 150+ million tweets related to COVID-19 for open research*. Tech. rep. Version 4.0. This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates.

Baroni, Marco, 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*. Vol. 1. Baltimore, USA, pp. 238–247.

Bartunov, Sergey, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov (2016). “Breaking sticks and ambiguities with adaptive skip-gram”. In: *Artificial Intelligence and Statistics*, pp. 130–138.

Basile, Pierpaolo, Annalina Caputo, and Giovanni Semeraro (2014). “Analysing Word Meaning over Time by Exploiting Temporal Random Indexing”. In: *The First Italian Conference on Computational Linguistics CLiC-it 2014*, p. 38.

Basile, Pierpaolo and Barbara McGillivray (2018). “Exploiting the Web for Semantic Change Detection”. In: *Discovery Science*. Ed. by Soldatova, Larisa, Vanschoren, Joaquin, Papadopoulos, George, and Ceci, Michelangelo. Springer International Publishing: Cham, pp. 194–208.

Bender, Emily M. and Alexander Koller (2020). “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Online, pp. 5185–5198.

Bengio, Yoshua, Rejean Ducharme, and Pascal Vincent (2003). “A Neural probabilistic language model”. In: *Journal of Machine Learning Research* vol. 3, pp. 1137–1155.

Berberich, Klaus, Srikanta J Bedathur, Mauro Sozio, and Gerhard Weikum (2009). “Bridging the Terminology Gap in Web Archive Search.” In: *WebDB*.

Bickel, Balthasar and Raymond Hickey (2017). “Areas and universals”. In: *Cambridge Handbooks in Language and Linguistics*, pp. 40–55.

Bizzoni, Yuri, Marius Mosbach, Dietrich Klakow, and Stefania Degaetano-Ortlieb (2019). “Some steps towards the generation of diachronic WordNets”. In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Linköping University Electronic Press: Turku, Finland, pp. 55–64.

Blank, Andreas (1999). “Why do new meanings occur? A cognitive typology of the motivations for lexical Semantic change”. In: *Historical Semantics and Cognition*. Mouton de Gruyter: Berlin/New York, pp. 61–90.

Blank, Andreas and Peter Koch (1999). *Historical semantics and cognition*. Vol. 13. Walter de Gruyter.

Blei, David M and John D Lafferty (2006). “Dynamic topic models”. In: *Proceedings of the 23rd international conference on Machine learning*. ACM, pp. 113–120.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent Dirichlet allocation”. In: *the Journal of Machine Learning Research* vol. 3, pp. 993–1022.

Bloomfield, Leonard (1933). *Language*. Allen & Unwin.

Bochkarev, Vladimir, Valery Solovyev, and Sören Wichmann (2014). “Universals versus historical contingencies in lexical evolution”. In: *Journal of The Royal Society Interface* vol. 11, no. 101, p. 20140841.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). “Enriching Word Vectors with Subword Information”. In: *Transactions of the Association of Computational Linguistics* vol. 5, pp. 135–146.

Boleda, Gemma and Katrin Erk (2015). “Distributional semantic features as semantic primitives—or not”. In: *2015 Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium Series*.

Borkowska, Paulina and Grzegorz Kleparski (2007). “It befalls words to fall down: pejoration as a type of semantic change”. In: *Studia Anglica Resoviensia*. Vol. 47(4), pp. 33–50.

Boser, Bernhard E, Isabelle M Guyon, and Vladimir N Vapnik (1992). “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, pp. 144–152.

Bréal, Michel (1899). *Essai de sémantique (2nd ed.)* Hachette: Paris.

Bullinaria, John A and Joseph P Levy (2007). “Extracting semantic representations from word co-occurrence statistics: A computational study”. In: *Behavior research methods* vol. 39, no. 3, pp. 510–526.

Chen, Dawn, Joshua Peterson, and Thomas Griffiths (2017). “Evaluating vector-space models of analogy. arXiv preprint”. In: *arXiv preprint arXiv:1705.04416*.

Chen, Zhigang, Wei Lin, Qian Chen, Xiaoping Chen, Si Wei, Hui Jiang, and Xiaodan Zhu (2015). “Revisiting Word Embedding for Contrasting Meaning”. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics: Beijing, China, pp. 106–115.

Chiu, Billy, Anna Korhonen, and Sampo Pyysalo (2016). “Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics: Berlin, Germany, pp. 1–6.

Choi, Hyunyoung and Hal Varian (2012). “Predicting the present with Google Trends”. In: *Economic Record* vol. 88, no. s1, pp. 2–9.

Chollet, Francois et al. (2015). *Keras*.

Chomsky, Noam (1965). *Aspects of the Theory of Syntax*. MIT press.

Church, Kenneth Ward and Patrick Hanks (1989). “Word Association Norms, Mutual Information, and Lexicography”. In: *27th Annual Meeting of the Association for Computational Linguistics*.

Cook, Paul, 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*. Dublin City University and Association for Computational Linguistics: Dublin, Ireland, pp. 1624–1635.

Cook, Paul, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin (2013). “A lexicographic appraisal of an automatic approach for detecting new word senses”. In: *Proceedings of eLex*.

Corney, David, Dyaa Albakour, Miguel Martinez, and Samir Moussa (2016). "What do a Million News Articles Look like?" In: *Proceedings of the First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016), Padua, Italy, March 20, 2016.* Pp. 42–47.

Croicu, Mihai and Ralph Sundberg (2015). "UCDP georeferenced event dataset codebook version 4.0". In: *Journal of Peace Research* vol. 50, no. 4, pp. 523–532.

Daniel, Michael and Nina Dobrushina (2016). *Two centuries in twenty words (in Russian).* NRU HSE.

Davies, Mark (2012). "Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English". In: *Corpora* vol. 7, no. 2, pp. 121–157.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman (1990). "Indexing by latent semantic analysis". In: *Journal of the American society for information science* vol. 41, no. 6, p. 391.

Del Tredici, Marco, Raquel Fernández, and Gemma Boleda (2019). "Short-Term Meaning Shift: A Distributional Exploration". 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).* Association for Computational Linguistics: Minneapolis, Minnesota, pp. 2069–2075.

Desagulier, Guillaume (2017). *Can word vectors help corpus linguists?* Tech. rep. Université Paris Nanterre.

Devlin, Jacob, 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).* Association for Computational Linguistics: Minneapolis, Minnesota, pp. 4171–4186.

Di Carlo, Valerio, Federico Bianchi, and Matteo Palmonari (2019). "Training temporal word embeddings with a compass". In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 33, pp. 6326–6334.

Drozd, Aleksandr, Anna Gladkova, and Satoshi Matsuoka (2016). "Word Embeddings, Analogies, and Machine Learning: Beyond king - man + woman = queen". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers.* The COLING 2016 Organizing Committee: Osaka, Japan, pp. 3519–3530.

Dubossarsky, Haim (2018). "Semantic change at large: A computational approach for semantic change research". PhD thesis. Ph. D. thesis, Hebrew University of Jerusalem, Edmond and Lily Safra Center for Brain Sciences.

Dubossarsky, Haim, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg (2019). "Time-Out: Temporal Referencing for Robust Modeling of Lexical Semantic Change". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics: Florence, Italy.

Dubossarsky, Haim, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman (2015). “A bottom up approach to category mapping and meaning change”. In: *Proceedings of the NetWordS 2015 Word Knowledge and Word Usage.* Pisa, Italy, pp. 66–70.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman (2016). “Verbs change more than nouns: a bottom-up computational approach to semantic change”. In: *Lingue e linguaggio* vol. 15, no. 1, pp. 7–28.

Dubossarsky, Haim, Daphna Weinshall, and Eitan Grossman (2017). “Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics: Copenhagen, Denmark, pp. 1147–1156.

Eck, Kristine (2005). *A beginner’s guide to conflict data: finding and using the right dataset.* Tech. rep. Department of Peace and Conflict Research, Uppsala University.

Eger, Steffen and Alexander Mehler (2016). “On the Linearity of Semantic Change: Investigating Meaning Variation via Dynamic Graph Models”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* Association for Computational Linguistics: Berlin, Germany, pp. 52–58.

Ehrmann, Maud, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano, and Roberto Navigli (2014). “Representing Multilingual Data as Linked Data: the Case of BabelNet 2.0”. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14).* European Language Resources Association (ELRA): Reykjavik, Iceland, pp. 401–408.

Ethayarajh, Kawin (2019). “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics: Hong Kong, China, pp. 55–65.

Fang, Meng and Trevor Cohn (2016). “Learning when to trust distant supervision: An application to low-resource POS tagging using cross-lingual projection”. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning.* Association for Computational Linguistics: Berlin, Germany, pp. 178–186.

Fares, Murhaf, Andrey Kutuzov, Stephan Oepen, and Erik Velldal (2017). “Word vectors, reuse, and replicability: Towards a community repository of large-text resources”. In: *Proceedings of the 21st Nordic Conference on Computational Linguistics.* Association for Computational Linguistics: Gothenburg, Sweden, pp. 271–276.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith (2015). “Retrofitting Word Vectors to Semantic Lexicons”. In: *Proceedings of the 2015 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics: Denver, Colorado, pp. 1606–1615.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer (2016). “Problems With Evaluation of Word Embeddings Using Word Similarity Tasks”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics: Berlin, Germany, pp. 30–35.

Firth, John (1935). “The Technique of Semantics.” In: *Transactions of the philological society* vol. 34, no. 1, pp. 36–73.

Firth, John (1957). *A synopsis of linguistic theory, 1930-1955*. Blackwell.

Fomin, Vadim, Daria Bakshandaeva, Julia Rodina, and Andrey Kutuzov (2019). “Tracing cultural diachronic semantic shifts in Russian using word embeddings: test sets and baselines”. In: *Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Dialog conference*, pp. 203–218.

Frermann, Lea and Mirella Lapata (2016). “A Bayesian Model of Diachronic Meaning Change”. In: *Transactions of the Association of Computational Linguistics* vol. 4, pp. 31–45.

Frey, Brendan and Delbert Dueck (2007). “Clustering by Passing Messages Between Data Points”. In: *Science* vol. 315, no. 5814, pp. 972–976.

Gábor, Kata, Haifa Zargayouna, Isabelle Tellier, Davide Buscaldi, and Thierry Charnois (2017). “Exploring Vector Spaces for Semantic Relations”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Copenhagen, Denmark, pp. 1815–1824.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou (2018). “Word embeddings quantify 100 years of gender and ethnic stereotypes”. In: *Proceedings of the National Academy of Sciences* vol. 115, no. 16, E3635–E3644.

Geeraerts, Dirk (1997). *Diachronic prototype semantics: A contribution to historical lexicology*. Clarendon Press: Oxford.

Gerow, Aaron and Khurshid Ahmad (2012). “Diachronic Variation in Grammatical Relations”. In: *Proceedings of COLING 2012: Posters*. The COLING 2012 Organizing Committee: Mumbai, India, pp. 381–390.

Giulianelli, Mario, Marco Del Tredici, and Raquel Fernández (2020). “Analysing Lexical Semantic Change with Contextualised Word Representations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Online, pp. 3960–3973.

Gladkova, Anna, Aleksandr Drozd, and Satoshi Matsuoka (2016). “Analogy-Based Detection of Morphological and Semantic Relations with Word Embeddings: What Works and What Doesn’t”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; student research workshop*. ACL: San Diego, California, pp. 47–54.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand (2002). “Armed conflict 1946-2001: A new dataset”. In: *Journal of peace research* vol. 39, no. 5, pp. 615–637.

Goldberg, Yoav (2017). “Neural network methods for natural language processing”. In: *Synthesis Lectures on Human Language Technologies* vol. 10, no. 1, pp. 1–309.

Golub, G. H. and C. Reinsch (1970). “Singular value decomposition and least squares solutions”. In: *Numerische Mathematik* vol. 14, no. 5, pp. 403–420.

Gonen, Hila, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg (2020). “Simple, Interpretable and Stable Method for Detecting Words with Usage Change across Corpora”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics: Online, pp. 538–555.

Gower, John C, Garmt B Dijksterhuis, et al. (2004). *Procrustes problems.* Vol. 30. Oxford University Press on Demand.

Greenawald, B., Y. Liu, G. Wert, M. A. Boni, and D. E. Brown (2018). “A comparison of language-dependent and language-independent models for violence prediction”. In: *2018 Systems and Information Engineering Design Symposium (SIEDS)*, pp. 260–265.

Gries, Stefan Th. (1999). “Particle movement: a cognitive and functional approach”. In: *Cognitive Linguistics* vol. 10, pp. 105–145.

Grzega, Joachim and Marion Schoener (2007). “English and General Historical Lexicology”. In: *Eichstätt-Ingolstadt: Katholische Universität.*

Gulordava, Kristina 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.* Association for Computational Linguistics: Edinburgh, UK, pp. 67–71.

Hamilton, William, Kevin Clark, Jure Leskovec, and Dan Jurafsky (2016). “Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Austin, Texas, pp. 595–605.

Hamilton, William, Jure Leskovec, and Dan Jurafsky (2016a). “Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics: Austin, Texas, pp. 2116–2121.

Hamilton, William, Jure Leskovec, and Dan Jurafsky (2016b). “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).* Association for Computational Linguistics: Berlin, Germany, pp. 1489–1501.

Harris, Zellig S (1954). “Distributional structure”. In: *Word* vol. 10, no. 2-3, pp. 146–162.

Hätty, Anna, Dominik Schlechtweg, and Sabine Schulte im Walde (2019). “SURel: A Gold Standard for Incorporating Meaning Shifts into Term

Extraction". In: *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*. Association for Computational Linguistics: Minneapolis, Minnesota, pp. 1–8.

Hellrich, Johannes, Sven Buechel, and Udo Hahn (2018). "JeSemE: Interleaving Semantics and Emotions in a Web Service for the Exploration of Language Change Phenomena". In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics: Santa Fe, New Mexico, pp. 10–14.

Hewitt, John and Christopher D. Manning (2019). "A Structural Probe for Finding Syntax in Word 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)*. Association for Computational Linguistics: Minneapolis, Minnesota, pp. 4129–4138.

Heyer, Gerhard, Florian Holz, and Sven Teresniak (2009). "Change of Topics over Time-Tracking Topics by their Change of Meaning." In: *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval (KDIR)* vol. 9, pp. 223–228.

Heyer, Gerhard, Cathleen Kantner, Andreas Niekler, Max Overbeck, and Gregor Wiedemann (2016). "Modeling the dynamics of domain specific terminology in diachronic corpora". In: *Proceedings of the 12th International conference on Terminology and Knowledge Engineering (TKE 2016)*.

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* vol. 41, no. 4, pp. 665–695.

Hilpert, Martin (2008). *Germanic future constructions: A usage-based approach to language change*. Benjamins: Amsterdam, Netherlands.

Hilpert, Martin and Stefan Th. Gries (2009). "Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition". In: *Literary and Linguistic Computing* vol. 24, no. 4, pp. 385–401.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-term Memory". In: *Neural computation* vol. 9, pp. 1735–80.

Hock, Hans Henrich and Brian D Joseph (2019). *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Walter de Gruyter GmbH & Co KG.

Hofmann, Thomas (1999). "Probabilistic Latent Semantic Analysis". In: *Proceedings of the Fifteenth conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden, pp. 289–296.

Hu, Minqing and Bing Liu (2004). "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 168–177.

Hu, Renfen, Shen Li, and Shichen Liang (2019). "Diachronic Sense Modeling with Deep Contextualized Word Embeddings: An Ecological View". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Florence, Italy, pp. 3899–3908.

Hürriyetoğlu, Ali, Vanni Zavarella, Hristo Tanev, Erdem Yörük, Ali Safaya, and Osman Mutlu (2020). “Automated Extraction of Socio-political Events from News (AESPEN): Workshop and Shared Task Report”. English. In: *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. European Language Resources Association (ELRA): Marseille, France, pp. 1–6.

Jaccard, Paul (1901). *Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines*. Rouge.

Jatowt, Adam, Ricardo Campos, Sourav S. Bhowmick, Nina Tahmasebi, and Antoine Doucet (2018). “Every Word Has Its History: Interactive Exploration and Visualization of Word Sense Evolution”. In: *CIKM ’18 Proceedings of the 27th ACM International Conference on Information and Knowledge Management, October 22 - 26, 2018, Torino, Italy*. ACM.

Jatowt, Adam and Kevin Duh (2014). “A framework for analyzing semantic change of words across time”. In: *IEEE/ACM Joint Conference on Digital Libraries*. IEEE, pp. 229–238.

Ji, Heng and Ralph Grishman (2008). “Refining Event Extraction through Cross-Document Inference”. In: *Proceedings of the 2008 Conference of the Association for Computational linguistics: Human Language Technologies*. Association for Computational Linguistics: Columbus, Ohio, pp. 254–262.

Johannessen, Janne Bondi, Kristin Hagen, André Lynum, and Anders Nøklestad (2012). “OBT+ stat. A combined rule-based and statistical tagger”. In: *I Andersen, Gisle (red.): Exploring Newspaper Language. Corpus compilation and research based on the Norwegian Newspaper Corpus*. John Benjamins Publishing Company.

Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2019). “Billion-scale similarity search with GPUs”. In: *IEEE Transactions on Big Data*.

Johnson, Melvin et al. (2017). “Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation”. In: *Transactions of the Association for Computational Linguistics* vol. 5, pp. 339–351.

Jones, Karen Sparck (1972). “A statistical interpretation of term specificity and its application in retrieval”. In: *Journal of documentation*.

Juola, Patrick (2003). “The Time Course of Language Change”. In: *Computers and the Humanities* vol. 37, no. 1, pp. 77–96.

Jurgens, David, Saif Mohammad, Peter Turney, and Keith Holyoak (2012). “SemEval-2012 Task 2: Measuring Degrees of Relational Similarity”. In: **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*. Association for Computational Linguistics: Montrèal, Canada, pp. 356–364.

Jurgens, David and Keith Stevens (2009). *Event Detection in Blogs using Temporal Random Indexing*. Borovets, Bulgaria.

K, Vani, Simone Mellace, and Alessandro Antonucci (2020). “Temporal Embeddings and Transformer Models for Narrative Text Understanding”. In: *Third International Workshop on Narrative Extraction from Texts (Text2Story*

2020) held in conjunction with the 42nd European Conference on Information Retrieval (ECIR).

Kaiser, Jens, Dominik Schlechtweg, Sean Papay, and Sabine Schulte im Walde (2020). "IMS at SemEval-2020 Task 1: How low can you go? Dimensionality in Lexical Semantic Change Detection". In: *arXiv preprint arXiv:2008.03164 (to appear in Proceedings of the 14th International Workshop on Semantic Evaluation).* Association for Computational Linguistics: Barcelona, Spain.

Kaji, Nobuhiro and Hayato Kobayashi (2017). "Incremental Skip-gram Model with Negative Sampling". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics: Copenhagen, Denmark, pp. 363–371.

Kanerva, Pentti, Jan Kristofersson, and Anders Holst (2000). "Random indexing of text samples for latent semantic analysis". In: *Proceedings of the 22nd annual conference of the cognitive science society.* Vol. 1036. Mahwah, USA.

Kendall, Maurice George (1948). *Rank correlation methods.* Griffin.

Kerremans, D., S. Stegmayr, and H.-J. Schmid (2010). "The NeoCrawler: Identifying and retrieving neologisms from the Internet and monitoring ongoing change". In: *Current methods in historical semantics.* Ed. by Allan, K. and Robinson, J. A. De Gruyter Mouton, pp. 130–160.

Kilgarriff, Adam (1997). "I don't believe in word senses". In: *Computers and the Humanities* vol. 31, no. 2, pp. 91–113.

Kim, Yoon, 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.* Association for Computational Linguistics: Baltimore, MD, USA, pp. 61–65.

Kulkarni, Vivek, 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.* Florence, Italy, pp. 625–635.

Kutuzov, Andrey, Vadim Fomin, Vladislav Mikhailov, and Julia Rodina (2020). "ShiftRy: web service for diachronic analysis of Russian news". In: *Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Dialog conference*, pp. 485–501.

Kutuzov, Andrey and Mario Giulianelli (2020). "UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection". In: *arXiv preprint arXiv:2005.00050 (to appear in Proceedings of the 14th International Workshop on Semantic Evaluation).* Association for Computational Linguistics: Barcelona, Spain.

Kutuzov, Andrey, Mikhail Kopotev, Tatyana Sviridenko, and Lyubov Ivanova (2016). "Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints". In: *Ninth Workshop on Building and Using Comparable Corpora at the Language Resources and Evaluation Conference (LREC).*

Kutuzov, Andrey and Maria Kunilovskaya (2017). "Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and

Russian National Corpus". In: *Analysis of Images, Social Networks and Texts*. Ed. by Aalst, Wil M.P. van der et al. Springer International Publishing: Cham, pp. 47–58.

Kutuzov, Andrey and Elizaveta Kuzmenko (2015). "Comparing neural lexical models of a classic national corpus and a web corpus: The case for Russian". In: *Lecture Notes in Computer Science* vol. 9041, pp. 47–58.

Kutuzov, Andrey and Elizaveta Kuzmenko (2016). "Cross-lingual Trends Detection for Named Entities in News Texts with Dynamic Neural Embedding Models". In: *First International Workshop on Recent Trends in News Information Retrieval co-located with 38th European Conference on Information Retrieval (ECIR 2016).* Technical University of Aachen, pp. 27–32.

Kutuzov, Andrey and Elizaveta Kuzmenko (2019). "To Lemmatize or Not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation". In: *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing.* Linköping University Electronic Press: Turku, Finland, pp. 22–28.

Kutuzov, Andrey, Elizaveta Kuzmenko, and Anna Marakasova (2016). "Exploration of register-dependent lexical semantics using word embeddings". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH).* The COLING 2016 Organizing Committee: Osaka, Japan, pp. 26–34.

Kutuzov, Andrey, 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.* Association for Computational Linguistics: Santa Fe, New Mexico, USA, pp. 1384–1397.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2017a). "Temporal dynamics of semantic relations in word embeddings: an application to predicting armed conflict participants". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics: Copenhagen, Denmark, pp. 1824–1829.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2017b). "Tracing armed conflicts with diachronic word embedding models". In: *Proceedings of the Events and Stories in the News Workshop.* Association for Computational Linguistics: Vancouver, Canada, pp. 31–36.

Kutuzov, Andrey, Erik Velldal, and Lilja Øvrelid (2019). "One-to-X Analogical Reasoning on Word Embeddings: a Case for Diachronic Armed Conflict Prediction from News Texts". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change.* Association for Computational Linguistics: Florence, Italy, pp. 196–201.

Landauer, Thomas K. and Susan T. Dumais (1997). "A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge". In: *Psychological Review* vol. 104, no. 2, pp. 211–240.

Lau, Jey Han, 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. Association for Computational Linguistics: Avignon, France, pp. 591–601.

Le, Quoc and Tomas Mikolov (2014). “Distributed representations of sentences and documents”. In: *International Conference on Machine Learning*, pp. 1188–1196.

Levy, Omer and Yoav Goldberg (2014). “Neural word embedding as implicit matrix factorization”. In: *Advances in Neural Information Processing Systems*, pp. 2177–2185.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). “Improving distributional similarity with lessons learned from word embeddings”. In: *Transactions of the Association for Computational Linguistics* vol. 3, pp. 211–225.

Liao, Xuanyi and Guang Cheng (2016). “Analysing the Semantic Change Based on Word Embedding”. In: *Natural Language Understanding and Intelligent Applications*. Ed. by Lin, Chin-Yew, Xue, Nianwen, Zhao, Dongyan, Huang, Xuanjing, and Feng, Yansong. Springer International Publishing: Cham, pp. 213–223.

Lijffijt, Jefrey, Tanja Säily, and Terttu Nevalainen (2012). “CEECing the baseline: Lexical stability and significant change in a historical corpus”. In: *Studies in Variation, Contacts and Change in English*. Vol. 10. Research Unit for Variation, Contacts and Change in English (VARIENG).

Lin, Jianhua (1991). “Divergence Measures Based on the Shannon Entropy”. In: *IEEE Transactions on Information theory* vol. 37, no. 1, pp. 145–151.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, eds. (2019). *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics: Florence, Italy.

Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang (2015). “Learning context-sensitive word embeddings with neural tensor skip-gram model”. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Liu, Qianchu, Diana McCarthy, Ivan Vulić, and Anna Korhonen (2019). “Investigating Cross-Lingual Alignment Methods for Contextualized Embeddings with Token-Level Evaluation”. In: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics: Hong Kong, China, pp. 33–43.

Logacheva, Varvara et al. (2020). “Word Sense Disambiguation for 158 Languages using Word Embeddings Only”. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association: Marseille, France, pp. 5943–5952.

Loukachevitch, Natalia and Anatolii Levchik (2016). “Creating a General Russian Sentiment Lexicon”. In: *Proceedings of Language Resources and Evaluation Conference (LREC-2016)*, pp. 1171–1176.

Loureiro, Daniel and Alipio Jorge (2019). “Language Modelling Makes Sense: Propagating Representations through WordNet for Full-Coverage Word Sense Disambiguation”. In: *Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics. Association for Computational Linguistics: Florence, Italy, pp. 5682–5691.

Manion, Steve Lawrence (2014). “Unsupervised Knowledge-based Word Sense Disambiguation: Exploration & Evaluation of Semantic Subgraphs”. PhD thesis. University of Canterbury. Department of Mathematics & Statistics.

Manning, Christopher D. (2015). “Computational Linguistics and Deep Learning”. In: *Computational Linguistics* vol. 41, no. 4, pp. 701–707.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. In: *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60.

Martin Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org.

Martinc, Matej, Petra Kralj Novak, and Senja Pollak (2020). “Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift”. English. In: *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association: Marseille, France, pp. 4811–4819.

Martinc, Matej, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova (2020). “Capturing Evolution in Word Usage: Just Add More Clusters?” In: *Companion Proceedings of the International World Wide Web Conference*, pp. 20–24.

McAuley, Julian John and Jure Leskovec (2013). “From Amateurs to Connoisseurs: Modeling the Evolution of User Expertise Through Online Reviews”. In: *Proceedings of the 22Nd International Conference on World Wide Web*. WWW ’13. ACM: Rio de Janeiro, Brazil, pp. 897–908.

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher (2017). “Learned in translation: Contextualized word vectors”. In: *Advances in Neural Information Processing Systems*, pp. 6294–6305.

McEnery, Tony, Vaclav Brezina, and Helen Baker (2019). “Usage Fluctuation Analysis: A new way of analysing shifts in historical discourse”. In: *International Journal of Corpus Linguistics* vol. 24, no. 4, pp. 413–444.

Melamud, Oren, Jacob Goldberger, and Ido Dagan (2016). “context2vec: Learning Generic Context Embedding with Bidirectional LSTM”. In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics: Berlin, Germany, pp. 51–61.

Michel, Jean-Baptiste et al. (2011). “Quantitative Analysis of Culture Using Millions of Digitized Books”. In: *Science* vol. 331, no. 6014, pp. 176–182.

Mihalcea, Rada, Timothy Chklovski, and Adam Kilgarriff (2004). “The Senseval-3 English lexical sample task”. In: *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics: Barcelona, Spain, pp. 25–28.

Mihalcea, Rada and Vivi Nastase (2012). “Word Epoch Disambiguation: Finding How Words Change Over Time”. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics: Jeju Island, Korea, pp. 259–263.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781*.

Mikolov, Tomas, Quoc Le, and Ilya Sutskever (2013). “Exploiting similarities among languages for machine translation”. arXiv preprint arXiv:1309.4168.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean (2013). “Distributed Representations of Words and Phrases and their Compositionality”. In: *Advances in Neural Information Processing Systems 26*, pp. 3111–3119.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). “Linguistic Regularities in Continuous Space Word Representations”. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics: Atlanta, Georgia, pp. 746–751.

Miller, George (1995). “WordNet: a lexical database for English”. In: *Communications of the ACM* vol. 38, no. 11, pp. 39–41.

Mitra, Sunny, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal (2014). “That’s 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).* Association for Computational Linguistics: Baltimore, Maryland, pp. 1020–1029.

Mueller, Hannes and Christofer Rauh (2017). “Reading Between the Lines: Prediction of Political Violence Using Newspaper Text”. In: *American Political Science Review*, pp. 1–18.

Nölle, Jonas, Stefan Hartmann, and Peeter Tinits (2020). “Language evolution research in the year 2020”. In: *Language Dynamics and Change* vol. 10, no. 1, pp. 3–26.

Okurowski, Mary Ellen (1993). “Domain and Language Evaluation Results”. In: *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993.*

Ono, Masataka, Makoto Miwa, and Yutaka Sasaki (2015). “Word Embedding-based Antonym Detection using Thesauri and Distributional Information”. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Association for Computational Linguistics: Denver, Colorado, pp. 984–989.

Orlikowski, Matthias, Matthias Hartung, and Philipp Cimiano (2018). “Learning Diachronic Analogies to Analyze Concept Change”. In: *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* Association for Computational Linguistics: Santa Fe, New Mexico, pp. 1–11.

Osgood, Charles Egerton, George J Suci, and Percy H Tannenbaum (1964). *The measurement of meaning.* University of Illinois Press.

Parker, Robert, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda (2011). *English Gigaword Fifth Edition LDC2011T07.* Tech. rep. Technical Report. Linguistic Data Consortium, Philadelphia.

Paszke, Adam et al. (2019). “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems*, pp. 8026–8037.

Pearson, Karl (1901). “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* vol. 2, no. 11, pp. 559–572.

Pelevina, Maria, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko (2016). “Making Sense of Word Embeddings”. In: *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics: Berlin, Germany, pp. 174–183.

Peng, Hao, Jianxin Li, Yangqiu Song, and Yaopeng Liu (2017). “Incrementally Learning the Hierarchical Softmax Function for Neural Language Models”. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, California USA, pp. 3267–327.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). “GloVe: Global Vectors for Word Representation”. In: *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543.

Peters, Matthew, 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)*. Association for Computational Linguistics: New Orleans, Louisiana, pp. 2227–2237.

Peters, Matthew, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih (2018). “Dissecting Contextual Word Embeddings: Architecture and Representation”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Brussels, Belgium, pp. 1499–1509.

Pettersson, Therése and Kristine Eck (2018). “Organized violence, 1989–2017”. In: *Journal of Peace Research* vol. 55, no. 4, pp. 535–547.

Pilehvar, Mohammad Taher and Jose Camacho-Collados (2019). “WiC: the Word-in-Context Dataset for Evaluating Context-Sensitive Meaning 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)*. Association for Computational Linguistics: Minneapolis, Minnesota, pp. 1267–1273.

Popescu, Octavian and Carlo Strapparava (2015). “SemEval 2015, Task 7: Diachronic Text Evaluation”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics: Denver, Colorado, pp. 870–878.

Popper, Karl (1962). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Řehůřek, Radim (2011). “Scalability of semantic analysis in natural language processing”. PhD thesis. Masaryk University.

Řehůřek, Radim and Petr Sojka (2010). “Software Framework for Topic Modelling with Large Corpora”. English. In: *Proceedings of the Language Resources and*

Evaluation Conference (LREC) 2010 Workshop on New Challenges for NLP Frameworks. ELRA: Valletta, Malta, pp. 45–50.

Rissanen, Matti et al. (1993). "The Helsinki Corpus of English Texts". In: *Kyttö et. al*, pp. 73–81.

Rodina, Julia, Daria Bakshandaeva, Vadim Fomin, Andrey Kutuzov, Samia Touileb, and Erik Velldal (2019). "Measuring Diachronic Evolution of Evaluative Adjectives with Word Embeddings: the Case for English, Norwegian, and Russian". In: *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*. Association for Computational Linguistics: Florence, Italy, pp. 202–209.

Rodina, Julia and Andrey Kutuzov (2020). "RuSemShift: a dataset of historical lexical semantic change in Russian". In: *arXiv:2010.06436 (to appear in Proceedings of the 28th Conference on Computational Linguistics (COLING-2020))*. arXiv: `2010.06436 [cs.CL]`.

Rogers, Anna, Aleksandr Drozd, and Bofang Li (2017). "The (too Many) Problems of Analogical Reasoning with Word Vectors". In: *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*. Association for Computational Linguistics: Vancouver, Canada, pp. 135–148.

Rong, Xin (2014). "word2vec parameter learning explained". In: *arXiv preprint arXiv:1411.2738*.

Rosenfeld, Alex 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*. New Orleans, Louisiana, USA, pp. 474–484.

Rosin, Guy D., Eytan Adar, and Kira Radinsky (2017). "Learning Word Relatedness over Time". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Copenhagen, Denmark, pp. 1179–1189.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* vol. 20, pp. 53–65.

Rubenstein, Herbert and John B Goodenough (1965). "Contextual correlates of synonymy". In: *Communications of the ACM* vol. 8, no. 10, pp. 627–633.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). "A survey of cross-lingual word embedding models". In: *Journal of Artificial Intelligence Research* vol. 65, pp. 569–631.

Rudolph, Maja and David M Blei (2018). "Dynamic embeddings for language evolution". In: *Proceedings of the 2018 World Wide Web Conference*, pp. 1003–1011.

Ryskina, Maria, Ella Rabinovich, Taylor Berg-Kirkpatrick, David R Mortensen, and Yulia Tsvetkov (2020). "Where New Words Are Born: Distributional Semantic Analysis of Neologisms and Their Semantic Neighborhoods". In: *Proceedings of the Society for Computation in Linguistics* vol. 3, no. 1, pp. 43–52.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2009). "Semantic Density Analysis: Comparing Word Meaning Across Time and Phonetic Space". In:

Proceedings of the Workshop on Geometrical Models of Natural Language Semantics. Association for Computational Linguistics, pp. 104–111.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark (2011). “Tracing semantic change with latent semantic analysis”. In: *Current methods in historical semantics*, pp. 161–183.

Sahlgren, Magnus (2005). “An introduction to random indexing”. In: *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering*.

Sahlgren, Magnus (2008). “The distributional hypothesis”. In: *Italian Journal of Disability Studies* vol. 20, pp. 33–53.

Sandhaus, Evan (2008). “The New York Times annotated corpus overview”. In: *Linguistic Data Consortium, Philadelphia* vol. 6, no. 12, e26752.

Saussure, Ferdinand de (1916). *Course in general linguistics*. Duckworth.

Schlechtweg, Dominik, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde (2019). “A Wind of Change: Detecting and Evaluating Lexical Semantic Change across Times and Domains”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Florence, Italy, pp. 732–746.

Schlechtweg, Dominik, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi (2020). “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection”. In: *To appear in Proceedings of the 14th International Workshop on Semantic Evaluation*. Association for Computational Linguistics: Barcelona, Spain.

Schlechtweg, Dominik, Sabine Schulte im Walde, and Stefanie Eckmann (2018). “Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change”. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics: New Orleans, Louisiana, pp. 169–174.

Schuster, Tal, Ori Ram, Regina Barzilay, and Amir Globerson (2019). “Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing”. 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)*. Association for Computational Linguistics: Minneapolis, Minnesota, pp. 1599–1613.

Schütze, Hinrich (1998). “Automatic Word Sense Discrimination”. In: *Computational Linguistics* vol. 24, no. 1, pp. 97–123.

Shoemark, Philippa, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray (2019). “Room to Glo: A Systematic Comparison of Semantic Change Detection Approaches with Word Embeddings”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics: Hong Kong, China, pp. 66–76.

Singer, Joel David and Melvin Small (1972). *The wages of war, 1816-1965: A statistical handbook.* John Wiley & Sons.

Søgaard, Anders (2016). “Evaluating word embeddings with fMRI and eye-tracking”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP.* Association for Computational Linguistics: Berlin, Germany, pp. 116–121.

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

Stewart, Ian, Dustin Arendt, Eric Bell, and Svitlana Volkova (2017). “Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network”. In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017).*

Straka, Milan and Jana Straková (2017). “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe”. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Association for Computational Linguistics: Vancouver, Canada, pp. 88–99.

Sundberg, Ralph and Erik Melander (2013). “Introducing the UCDP georeferenced event dataset”. In: *Journal of Peace Research* vol. 50, no. 4, pp. 523–532.

Szymanski, Terrence (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).* Association for Computational Linguistics: Vancouver, Canada, pp. 448–453.

Tahmasebi, Nina (2013). “Models and algorithms for automatic detection of language evolution: towards finding and interpreting of content in long-term archives”. PhD thesis. Hannover: Gottfried Wilhelm Leibniz Universität Hannover.

Tahmasebi, Nina, Lars Borin, and Adam Jatowt (2018). “Survey of Computational Approaches to Lexical Semantic Change”. In: *Preprint at ArXiv 2018.*

Tahmasebi, Nina, Lars Borin, Adam Jatowt, and Yang Xu (2019). *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change.* Association for Computational Linguistics: Florence, Italy.

Tahmasebi, Nina, Gerhard Gossen, Nattiya Kanhabua, Helge Holzmann, and Thomas Risse (2012). “NEER: An Unsupervised Method for Named Entity Evolution Recognition”. In: *Proceedings of COLING 2012.* The COLING 2012 Organizing Committee: Mumbai, India, pp. 2553–2568.

Tahmasebi, Nina and Thomas Risse (2017a). “Finding Individual Word Sense Changes and their Delay in Appearance”. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017.* INCOMA Ltd.: Varna, Bulgaria, pp. 741–749.

Tahmasebi, Nina and Thomas Risse (2017b). *Word Sense Change Testset.* Tech. rep. This work has been funded in parts by the project "Towards a knowledge-based culturomics" supported by a framework grant from the Swedish Research Council (2012–2016; dnr 2012-5738). This work is also in parts funded by the European Research Council under Alexandria (ERC 339233) and the European Community's H2020 Program under SoBigData (RIA 654024).

Tang, Xuri (2018). "A state-of-the-art of semantic change computation". In: *Natural Language Engineering* vol. 24, no. 5, pp. 649–676.

Taylor, Wayne A (2000). *Change-point analysis: a powerful new tool for detecting changes.*

Tjong Kim Sang, Erik (2016). "Finding Rising and Falling Words". In: *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH).* The COLING 2016 Organizing Committee: Osaka, Japan, pp. 2–9.

Tollefsen, Andreas Forø, Håvard Strand, and Halvard Buhaug (2012). "PRIO-GRID: A unified spatial data structure". In: *Journal of Peace Research* vol. 49, no. 2, pp. 363–374.

Traugott, Elizabeth (1999). "The role of pragmatics in semantic change". In: *Pragmatics in 1998: Selected Papers from the 6th International Pragmatics Conference, vol.II.* International Pragmatics Association.

Traugott, Elizabeth (2017). "Semantic change". In: *Oxford Research Encyclopedias: Linguistics.*

Traugott, Elizabeth and Richard Dasher (2001). *Regularity in semantic change.* Cambridge University Press.

Tsakalidis, Adam, Marya Bazzi, Mihai Cucuringu, Pierpaolo Basile, and Barbara McGillivray (2019). "Mining the UK Web Archive for Semantic Change Detection". In: *Proceedings of Recent Advances in Natural Language Processing conference.*

Turney, Peter (2006). "Similarity of Semantic Relations". In: *Computational Linguistics* vol. 32, no. 3, pp. 379–416.

Turney, Peter, Patrick Pantel, et al. (2010). "From frequency to meaning: Vector space models of semantics". In: *Journal of artificial intelligence research* vol. 37, no. 1, pp. 141–188.

Ulčar, Matej and Marko Robnik-Šikonja (2020). "High Quality ELMo Embeddings for Seven Less-Resourced Languages". In: *arXiv preprint arXiv:1911.10049.*

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* vol. 9, no. 2579-2605, p. 85.

Van Durme, Benjamin and Ashwin Lall (2010). "Online Generation of Locality Sensitive Hash Signatures". In: *Proceedings of the Association for Computational Linguistics Conference (Short Papers).* Association for Computational Linguistics: Uppsala, Sweden, pp. 231–235.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Velldal, Erik (2011). "Random Indexing Re-Hashed". In: *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011).* Northern European Association for Language Technology (NEALT): Riga, Latvia, pp. 224–229.

Wallensteen, Peter (2013). *Peace Research: Theory and Practice*. Routledge.

Wang, Xuerui and Andrew McCallum (2006). "Topics over time: a non-Markov continuous-time model of topical trends". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 424–433.

Warren, Beatrice (1999). "Laws of thought, knowledge and lexical change". In: *Historical semantics and cognition*, pp. 215–34.

Wijaya, Derry Tanti 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*. ACM, pp. 35–40.

Wright, Quincy (1942). *A study of war*. University of Chicago Press.

Xu, Jin, Yubo Tao, Yuyu Yan, and Hai Lin (2019). "Exploring Evolution of Dynamic Networks via Diachronic Node Embeddings". In: *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1.

Xu, Yang and Charles Kemp (2015). "A Computational Evaluation of Two Laws of Semantic Change". In: *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2015)*.

Yaghoobzadeh, Yadollah, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze (2019). "Probing for Semantic Classes: Diagnosing the Meaning Content of Word Embeddings". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Florence, Italy, pp. 5740–5753.

Yaghoobzadeh, Yadollah and Hinrich Schütze (2016). "Intrinsic Subspace Evaluation of Word Embedding Representations". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics: Berlin, Germany, pp. 236–246.

Yao, Zijun, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong (2018). "Dynamic Word Embeddings for Evolving Semantic Discovery". In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. ACM: Marina Del Rey, CA, USA, pp. 673–681.

Yin, Zi, Vin Sachidananda, and Balaji Prabhakar (2018). "The global anchor method for quantifying linguistic shifts and domain adaptation". In: *Advances in Neural Information Processing Systems*, pp. 9433–9444.

Zalizniak, Anna (2018). "The Catalogue of Semantic Shifts: 20 Years Later". In: *Russian Journal of Linguistics* vol. 22, no. 4, pp. 770–787.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka (2015). "Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics: Beijing, China, pp. 645–655.

Zhang, Yating, Adam Jatowt, Sourav Bhowmick, and Katsumi Tanaka (2016). "The Past is Not a Foreign Country: Detecting Semantically Similar Terms

Across Time”. In: *IEEE Transactions on Knowledge and Data Engineering* vol. 28, no. 10, pp. 2793–2807.

Zipf, George Kingsley (1949). *Human behavior and the principle of least effort*. Addison-Wesley Press.

Zukov Gregoric, Andrej, Zhiyuan Luo, and Bartal Veyhe (2016). “IBC-C: A Dataset for Armed Conflict Analysis”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics: Berlin, Germany, pp. 374–379.