Research Master Thesis

Using Language Models for Analyzing Semantic Variation between Dutch Social Communities

Sanne Hoeken

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by:	Antske Fokkens and Pia Sommerauer
2^{nd} reader:	Angel Daza
Submitted:	June 30, 2022

Abstract

This thesis aims to exploit what information language models can provide about social semantic variation through a multidisciplinary approach. The increased impact of social media is said to result in growing political polarization through effects for which the concepts of *echo chamber* and *filter bubble* have been introduced. These worrying polarizing effects in society together with the hypothesis that they manifest in language, raise the need for more insights into semantic variation from a social perspective. In addition, the almost endless amount of user-generated web data, along with advanced language modelling techniques, provide potential research opportunities into these social phenomena.

Building on the novel field of Cognitive Sociolinguistics, expertise-informed hypotheses for a list of target words have been drawn up regarding conceptual variation between two Dutch online communities positioned at the extreme ends of the political spectrum. Several studies have already used language models for the analysis of diachronic semantic shifts, i.e. Lexical Semantic Change Detection (LSCD). However, much remained unclear about state-of-the-art contextualized language models as well as their application to the analysis of social variation.

The experiments in this study test two different LSCD methods in a social setting and establish the validity of the results on the target words through different control set-ups. The key findings show that word meaning representations created with a PPMI-based language model show hypothesized community-dependent semantic variation, while contextualized representations generated with the pre-trained contextualized language model BERTje do not. This outcome contradicts the high performance character of BERT-like models in the field of Natural Language Processing.

Experiments with control words and fine-grained analyses of PPMI representations indicated that different word interpretations by different political communities do not only apply to fundamentally political concepts, but manifests itself much more widely across the discourse. This points to a distinction between social and temporal semantic variation, which calls for a critical reconsideration of translating LSCD directly into the analysis of social semantic variation.

Declaration of Authorship

I, Sanne Naomi Hoeken, declare that this thesis, titled *Using Language Models for Analyzing Semantic Variation between Dutch Social Communities* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 30-06-2022

Signed: .

Acknowledgements

First of all, I would like to thank my supervisors Antske Fokkens and Pia Sommerauer. You put my research ideas into realistic perspective, supplemented and improved them. I am grateful for your cordial and accessible way of guidance. I would like to thank Mariken van der Velden for her expertise in the field of political communication. In addition to providing support in drafting and strengthening (the validity of) the hypotheses in this thesis, your enthusiasm motivated me even more.

This thesis was the completion of my education in which I was able to learn a lot from the CLTL staff members and fellow students from the Human Language Technology master. Specifically, I would like to express my gratitude to Piek Vossen, Lisa Beinborn and Levi Remijnse. Piek, I am in awe of the way in which you manage to combine an inhumane workload with such a human attitude. Lisa, I cannot deny that I found your lectures most interesting and enjoyable. Thank you for your honesty and critical but gentle attitude. Levi, I am thankful for our linguistic discussions about frames and implicatures (and the employee pass, and thus free coffee).

Finally, I am grateful for my friends, roommates and family. Writing a thesis is quite individual and stressful at times, but you guys provided me with company and relaxation when I needed it the most. In particular I would like to thank my father for the trust he gave me, my twin sister for leaving me during the whole thesis period to send pictures of thai curries from the other side of the world and my mother for pointing out that ‘acknowledgments’ should be written with an extra ‘e’.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Social Media and Political Polarization

Over the past decades, we have witnessed tremendous growth in digital technologies. This growth enforced the emergence of social media that play a dominant role in society today. More and more communication takes place online as various platforms such as Facebook and Reddit make it easy to make online connections. The great potential for finding like-minded people on social media platforms fuels the existence of online communities. A common factor underlying like-mindedness is political preference. Individuals with the same political convictions find each other online and exchange information and thoughts in their own circle. The opportunities to find agreeing beliefs are substantially greater with the advent of social media. The same applies to the tendencies to become closed off from deviant or opposing beliefs, which has been theorized with the introduced concepts of **echo chamber** and **filter bubble**. Sunstein (2001) introduced the term echo chamber, denoting the social phenomenon in which individuals acquire reinforced and more extreme political beliefs through greater and more selective exposure to consonant information and ideas on social media platforms. The selective aspect is what Pariser (2011) emphasized with his introduction of the filter bubble. The development of filtering algorithms implemented in social media technologies allows users to see personalized web content. This enhances the more selective information the user is exposed to, thereby limiting exposure to opposing or differing perspectives or beliefs. (Pariser, 2011). Movements towards the extremes of the political spectrum are increasing and so is the aversion to supporters at the other extreme of the spectrum. The latter can be referred to as **affective polarization** that seems to apply not only to two-party systems as in the US but also to multi-party systems as in the Netherlands. (Harteveld, 2021).

The notions of echo chambers and filter bubbles are not limited to scientific theories about social phenomena. Several political figures and policymakers have already publicly discussed the effects in their statements and plans. It was even President Obama who, in his Farewell Address in Chicago on January 10, 2017, mentioned the role of social media in the increasing political polarization. Obama described the filter bubble effect as the explanatory aspect:

> “And increasingly, we become so secure in our bubbles that we accept only information, whether true or not, that fits our opinions, instead of basing our opinions on the evidence that’s out there.” (Obama, 2017)

The notion of echo chamber is also discussed in last year's publication of Threat Assessment for the Netherlands by the NCTV (National Coordinator for Security and Counterterrorism), part of the Ministry of Justice and Security. With regard to certain right-wing extremist groups in the Netherlands, the possible cause of a terrorist threat is, among other things, attributed to online echo chambers:

> "The combination of this vulnerability with repeated immersion in online echo chambers can lead to radicalisation and even to the use of violence." (NCTV, 2021)

These examples show that the social phenomena under discussion are not only emerging in the current age of dominant online discourse, but also need attention and would benefit from more insight.

1.2 Cognitive Sociolinguistics

A fundamental element in conveying political beliefs is language. One of the leading scholars who shows how (the growing) political polarization is encoded in language, is cognitive linguist George Lakoff. As an American liberal, he sees the political polarization in America as a moral divide as political prescriptions are in fact visions of what is right or wrong. The divide is therefore the result of different views on what is right or wrong. In his research field of cognitive science, Lakoff investigates what determines our political behavior on the basis of frames: "mental structures that shape the way we see the world" (Lakoff, 2014). Each word activates a particular frame in our brain, making language not only the observable means to study frames, but also the weapon to give facts moral meaning. The relationship between language and political beliefs is a two-way street: the language chosen **reflects** a (political or moral) worldview and the chosen language **evokes** frames and thus the concepts that contribute to a (political or moral) worldview. Lakoff argues that American conservatives and progressives think with different frames, allowing words to have contestable meanings. An example of such a word, according to Lakoff, is a fundamental concept in politics: freedom (Lakoff et al., 2006).

With polarizing effects in society more topical than ever, and given the hypothesis that these effects manifest in our language, even at the level of word meaning, more insights into the social meaning of semantic variation would be valuable. Prior research mainly addressed two separate parts of this issue. First, cognitive linguistic research places semantics as their central component of analyses and studies the variation and change of meaning. Following the cognitive linguistic perspective, word meaning is grounded in physiological and social experience (Geeraerts et al., 2010). Second, sociolinguistic research investigates what linguistic variations and changes mean in a social context (Labov, 1966). The variationist or quantitative approach is a commonly used approach that correlates social factors with linguistic features such as phonological, morphological or syntactic properties. Research on *semantic* variation has remained limited in this field. Robinson (2012) explains the accepted methods in the field being mainly quantitative methods based on objective linguistic features. The fuzzy nature of word meaning makes extracting semantic features a complex matter not in accordance with the objective standards. Glynn (2010) argues that although extralinguistic context is integral to linguistic structure according to their perspective, sociolinguistic factors are often sidelined in cognitive linguistic research.

A handful of case-studies (e.g. Robinson (2010)) together with the contributions of Geeraerts et al. (1994) and Hasan (2009) show "that meaning variation and change is a complex socio-conceptual phenomenon" (Robinson, 2012). They demonstrate the added value of a new research area introduced by Geeraerts et al. (2010) as Cognitive Sociolinguistics.

1.3 Language Models

Cognitive Linguistics and Sociolinguistics overlap in methodological perspective focusing on empirical studies of language in use. Distributional semantics provide a way to approach word meaning empirically and lays the foundation for the computational modeling of word meaning, resulting in the development of so-called *language models*. Early language models were based on counting how often words appeared in each others context. Since the rise of Machine Learning techniques, predictive models have been shown to be capable of creating more qualitative and powerful word meaning representations than count-based representations. Especially the recently developed techniques in *contextualized* language models, which encompass deep neural network architectures with hundreds of millions of parameters, deliver impressive performance. These models, with BERT (Devlin et al., 2019) being one of the most popular examples, generate powerful context-specific word representations enabling the models to perform human-like accuracies on various Natural Language Processing tasks.

One of the tasks, that serves historical linguistic purposes, for which the use of computational technologies has already extensively been introduced is Lexical Semantic Change Detection (LSCD). Roughly speaking, the goal of this task is to detect semantic change by measuring difference between language model-generated meaning representations of the same words in different time-specific data. A wide variety of methods have already been developed for this task, although much is still unclear about the use of state-of-the-art contextualized language models.

Del Tredici and Fernández (2017) seem to be the first to show that the methods for the detection of diachronic semantic shifts, i.e. semantic variation over time, can also provide new insights into semantic variation in a social setting. With only a few studies so far, research into the use of LSCD methods for the analysis of synchronic variation, i.e. community dependent meaning difference, remains limited. Still, these recent studies indicate the potential for language models to study semantic variation from a socio-cognitive perspective.

1.4 Research Aims

With the role of social and mass media in society getting more and more important, the amount of user-generated web data is increasing rapidly. The almost endless amount of data, together with advanced language modeling techniques, provide potential research opportunities on sociolinguistic phenomena that would benefit from greater insight and understanding. This line of research into computational linguistic possibilities within the novel field of Cognitive Sociolinguistics has barely been touched upon. In this research I aim to answer the following question:

What information do language models provide about semantic variation between Dutch social communities?

The main-question is addressed by two sub-questions:

1. Do language models confirm cognitive-linguistically informed hypotheses about semantic variation in social settings?

2. What are the performance differences between transparent count-based and contextualized language models in Lexical Semantic Change Detection methods applied to a use case covering synchronic variation?

All in all, this research aims to investigate semantic variation in social settings using computational linguistic methods. It builds on 1) the cognitive sociolinguistic point of view that social factors are important for variation in word meaning, 2) the Distributional Hypothesis that states word meaning can be studied by looking at the contexts in which the word occurs and 3) the principle that methods for the analysis of diachronic semantic variation can be effectively translated into methods for the analysis of semantic variation across social communities.

1.5 Outline

The remainder of this thesis is structured as follows. The theoretical background of semantic variation and the different linguistic disciplines and computational methods that study semantic variation are discussed in Chapter 2. Chapter 3 explains the methodology used in this thesis to analyze semantic variation from a socio-cognitive perspective using computational techniques. The results of the performed experiments are presented in Chapter 4 and subsequently interpreted and discussed in Chapter 5. Finally, the conclusions that follow from the results, together with suggestions for further research, will be discussed in Chapter 6.

Chapter 2

Background and Related Work

This chapter discusses the relevant literature for the study of semantic variation between social communities. Specifically, the focus of this thesis is on the use of language models for these research purposes. Figure 2.1 shows the relationships between the topics and areas of research discussed. The central phenomenon of semantic variation will be introduced in Section 2.1. Several branches of theoretical linguistics study this phenomenon. Only the novel field of Cognitive Sociolinguistics (Section 2.1.1) addresses semantic variation from a socio-cognitive perspective. This branch, which originated from a combination of Cognitive Linguistics and Sociolinguistics, could benefit from methods used in Historical Linguistics (Section 2.1.2). Leveraging computational language models, which are discussed extensively in Section 2.2, for historical linguistic purposes resulted in the field of Lexical Semantic Change Detection (Section 2.3). As the focus of this thesis is methodological in nature, these methods will be discussed in detail in Section 2.3. Section 2.4 considers the first studies that apply LSCD methods to synchronic semantic variation, indicating the potential of using computational methods for Cognitive Sociolinguistic research issues. This potential underlies the research gap that I address in this thesis, which I finally introduce with a brief conclusion to the reviewed literature in Section 2.5.

2.1 Semantic Variation

Human language is a dynamic phenomenon that makes its variation and change a subject of several studies. Languages vary and change with regard to various features. These features can be clearly observable such as phonological and morphological features, but also less objective and fuzzier in nature, such as semantic features. This research focuses on the latter, concerning variation in word meaning. Given the interpretation that word meaning concerns the link between a lexical form and a particular concept, two directions of analysis can be distinguished here. Semasiology takes the lexical form as a starting point and looks at the concepts, or word meanings, it refers to. For example, the word ‘virus’ can refer to both a disease and a computer program. Onomasiology takes a particular concept as its starting point, and looks at the lexical forms by which this concept is referred to (Geeraerts, 1997). For example, the words ‘illness’, ‘disease’ and ‘virus’ could all refer to the same concept. This study concentrates on semasiological variation, i.e. how do the meanings of a particular word vary?

Change of meaning can manifest itself in different ways. Meanings can become

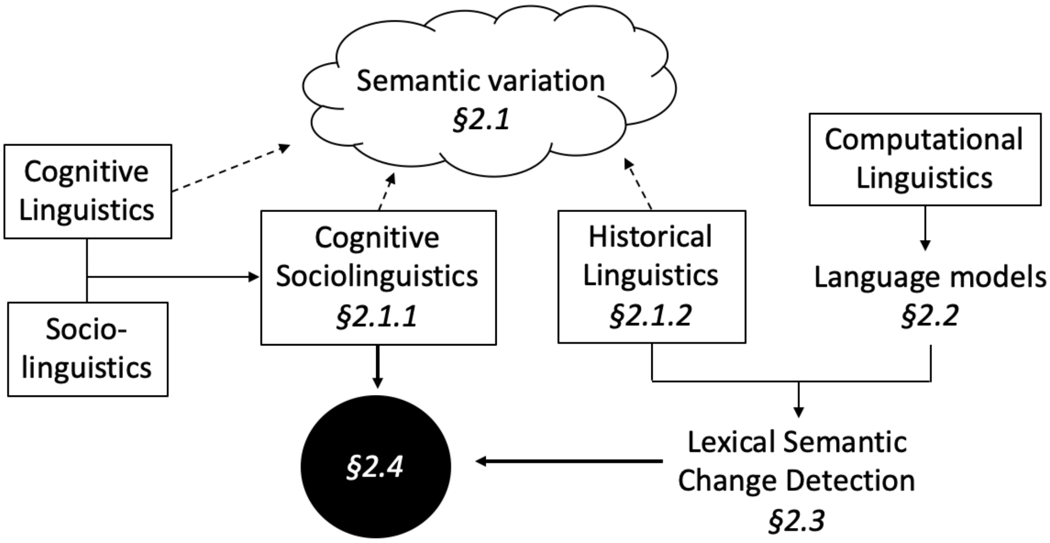


Figure 2.1: Background overview

broader, i.e. more general, or narrower, i.e. more specialized. Word meaning can also change at the level of connotation. Pejoration refers to a negative connotation shift and amelioration to a positive one (Ullmann, 1962).

2.1.1 Cognitive Sociolinguistics

As introduced in Chapter 1, the study of semantic variation in social context is a novel direction, bridging the gap between Cognitive Linguistics and Sociolinguistics. The first case studies show that this combination, resulting in the research field of Cognitive Sociolinguistics, can provide new insights. Glynn (2010) uses cognitive linguistic methods to show that specific meanings of 'annoy' (in terms of strength of emotion and presence of specific person as source of annoyance) correlate with sociolinguistic features. And starting from a sociolinguistic perspective, Robinson demonstrates differences between generations, socio-economic status and gender in the way speakers use the words 'awesome' (Robinson, 2010) and 'gay' (Robinson, 2012). The methods used in these case-studies studies are based on manual analyses. Scaling up is therefore complicated by the labor intensity of these type of methods. How computational techniques could be used for cognitive-sociolinguistic analyses is therefore a relevant issue. The use of computational techniques that enable somewhat similar linguistic analyses on a larger scale has already been applied more extensively in another branch of linguistics, i.e. Historical Linguistics.

2.1.2 Historical Linguistics

A related, older and more popular field of linguistics also concerned with semantic variation, is Historical Linguistics or Diachronic Linguistics. In order to gain a better understanding of changes in language over time, this research field also analyzes *semantic* shifts of language over time. Traditional methods, which did not yet involve computational techniques, addressed the analysis of semantic shifts by mainly relying on collocations (Traugott, 2017). The principle underlying these methods is that difference in expressions surrounding a particular word indicate different uses of the word, implying variation in meaning of the word. For example, the word ‘virus’ initially only appeared in contexts with words such as ‘bacteria’ and ‘illness’, and since a few decades ago this word also appears in contexts with words such as ‘computer’ and ‘software’. This shift of context words corresponds to the meaning change from a disease to also a computer program. This contextual approach conforms to the principles of Distributional Semantics, which also lays the foundations for state-of-the-art methods using computational techniques for the analysis of lexical semantic change. I will discuss these developments in the following sections.

2.2 Language Models

Distributional semantics provide a way to approach word meaning empirically. The idea that the meaning of a word can be defined as “its use in language” (Wittgenstein, 1953) resulted in the Distributional Hypothesis (Harris, 1954; Firth, 1957). This hypothesis states that if meaning is defined through use, then the meaning of a word can be defined by representing the contexts in which it occurs in a large corpus. A major advantage of this approach is that it allows creating so-called *language models* where word meaning is represented *numerically* resulting in high-dimensional vectors (where a vector is a list of numbers). In short, language models are models that assign probabilities to sequences of linguistic units. Words that occur in similar contexts are assigned similar probabilities, resulting in similar numerical representations. We can measure that such vectors are similar through e.g. their cosine distance.

Early language models were based on counting how often words appeared in each other’s context. In these count-based models, vector representations are a function of counts of neighboring words and can be derived from co-occurrence matrices (Jurafksy and Martin, 2021). I will explain this type of language models, and specifically the PPMI algorithm, in Section 2.2.1. Since the rise of Machine Learning techniques, predictive models have become dominant in the development of language models. If these models are provided with sufficient data, i.e. at least about hundreds of millions of words (Sahlgren and Lenci, 2016), they have proved capable of creating more qualitative and powerful word meaning representations than count-based representations. Prediction-based language models can be divided into two sub-directions: static word embedding models (Section 2.2.2) and contextual word embedding models (Section 2.2.3).

2.2.1 Count-based Language Models

The starting point of a count-based language model is a co-occurence matrix, in which both the number of rows and the number of columns are defined by the size of the vocabulary of a corpus. A row in the matrix represents the frequencies at which a word in the vocabulary occurs together with every other word in the vocabulary (the columns) in a context. The context usually considered is a window, where the size of the window reflects the number of words to the left and right of a target word, for which the co-occurences in a corpus are counted. The resulting co-occurence matrix is in itself ineffective to infer semantic relationships between words, since high-frequency but uninformative words such as ‘the’ and ‘a’ make meaningful discriminations between words impractical. The PPMI algorithm provides a solution to this limitation (Jurafksy and Martin, 2021).

Pointwise Mutual Information (PMI) between a word w and word c measures how often w occurs in a context with c in a corpus, compared to co-occurence of the words by chance. A negative PMI value would imply that two words co-occur less often than by chance. This is not reliable with a limited-sized corpus (Jurafksy and Martin, 2021). Therefore, the Positive Pointwise Mutual Information is commonly used, where all negative values are replaced by zero. Eliminating the negative values puts the emphasis of the representations on the positive word correlations, which appeared to improve the results (Levy et al., 2015). The mathematical details of the PPMI algorithm are presented in Section 3.3.1.

2.2.2 Static Word Embedding Models

Word vectors resulting from the previously described count-based models are sparse, i.e. a large proportion of the values are zeros. Computing dense word vectors, or embeddings, of several hundred dimensions rather than as large as the vocabulary has been found to result in more powerful meaning representations. One of the most popular type of methods to compute such representations are Word2Vec models (Mikolov et al., 2013), which include neural network inspired architectures. One of the most popular word embedding models is the Word2Vec model based on the Skip-Gram with Negative Sampling (SGNS) algorithm. The model is initialized with two random word embeddings for each word in the training vocabulary: a target word embedding and a context word embedding. During training, not only a positive example is used for a target word, i.e. a context word that actually co-occurs with the target word in the corpus, but also negative examples, i.e. random words with which the target word does not necessarily co-occur. The values of the embeddings are updated in such a way that the similarity between the target word and 'positive' context words is maximized and the similarity between the target word and 'negative' context words is minimized. The final set of embeddings learned share a multidimensional space in which words with similar meanings have similar vectors and therefore have a smaller distance from each other in the embedding space. The embeddings resulting from such a predictive method are *static*, i.e. for each word in the vocabulary one embedding is learned.

2.2.3 Contextualized Language Models

The methods described above to represent word meaning numerically are limited in their ability to model all natural properties of language. By taking a context window in which each context word is treated independently and equally with respect to the target word, the sequential nature of language can only be taken into account to a limited extend. In addition, because of the one-embedding-per-word principle, it is also not possible to capture the polysemic character of language. The advent of deep neural networks, and specifically Recurrent Neural Networks (RNNs), brought improvements to these limitations.

Language models with RNN architecture (e.g. Melamud et al. (2016); McCann et al. (2017)) are trained to predict the next word, one word at a time, in a sequence of words by including not only the original input of the current word but also the hidden representation of the previous timestep. In this way, context information is used sequentially from sentence start to current word to compute the probabilities. To also include the context information following the current word in the computation, a bidirectional RNN offers a solution. The sequence is not only processed from left to right, but also from right to left.

However, the longer the input sequence gets, the more distance information is lost during processing. A more complex RNN variant, the Long Short Term Memory (LSTM), mitigates this effect with so-called memory cells. The bidirectional LSTM architecture is the basis of ELMo (Peters et al., 2018), one of the first successful *contextualized* language models. Embeddings from Language Models (ELMo) are deep contextualized word representations. The representations are contextualized and deep as each token is represented by an embedding which is a function of the entire input sequence and also a function of all hidden layers in the deep neural network.

However, the memory cells in the LSTM architecture do not completely solve the memory problem with long sequences. ELMo was therefore followed by the even more successful BERT, Bidirectional Encoder Representations from Transformers. The fundamental building blocks of BERT are Transformers (Vaswani et al., 2017). The distinguishing components in these blocks, which further consist of standard neural network layers with some extra connections, are the self-attention layers. The main effect of these self-attention layers is that they allow the network to focus on various parts of a sequence, regardless of the distance, while word-for-word processing the sequence. BERT and related transformer-based models span hundreds of millions of parameters and are trained on data with billions of tokens.

The common training objective for transformer based language models is Masked Language Modeling (MLM). In the training corpus, a share of tokens is masked and the task is to predict which tokens have been there. In BERT-like models, the input sequences that are first tokenized, not into words, but into subwords are combined with embeddings that represent the position of the word in the sequence. The phase in which contextualized language models are trained in this way on a rich amount of data is the pre-training phase. Subsequently, the pre-trained embeddings can be used as a basis for further training, the fine-tune phase, on the same or another Natural Language Processing task. Transformer-based contextualized language models have been shown to outperform state-of-the-art models on many NLP tasks.

The original BERT was trained on English-language data, but after its success, BERT variants quickly followed for other languages, including Dutch. The Dutch variant is called BERTje, and just like the original BERT consists of 12 transformer blocks with 12 attention layers each and a hidden layer size of 768. The model therefore contains 100 million parameters, and concerns a vocabulary of 30k subwords. BERTje is, of course unlike the original BERT, pre-trained on a large and diverse amount of Dutch data with 2.4 billion tokens (de Vries et al., 2019).

2.3 Lexical Semantic Change Detection (LSCD)

Computational modeling of word meaning has been used with varying degrees of success for historical linguistic purposes, i.e. analyzing diachronic lexical semantic change (e.g. Tahmasebi (2013)). A common approach to this task of Lexical Semantic Change Detection (LSCD), schematized in Figure 2.2, involves 1) a corpus consisting of two or more subcorpora from different time periods, 2) the extraction of meaning representations for a set of target words from each subcorpus, and 3) the comparison of the representations to detect semantic change.

Within this general approach there is a great diversity of methods attempting to

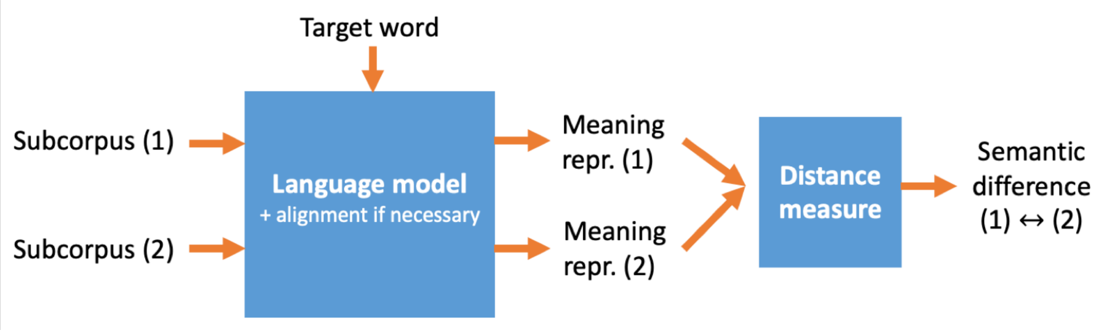


Figure 2.2: General schematization of a common approach in the field of Lexical Semantic Change Detection

solve the task. Two important components by which the different LSCD methods can be distinguished are: 1) the technique to derive word meaning representations from different subcorpora that can be fairly compared with each other and 2) the quantitative metric which is used to measure change between the representations. These components do not allow an exhaustive coverage of all existing LSCD methods. For example, existing LSCD methods based on models other than language models (e.g. topic (Fermann and Lapata, 2016) or graph-based (Mitra et al., 2015)) are not covered, due to the focus and scope of this thesis.

In the sections that follow, I will provide an overview of the leading methods for LSCD using the two components drawn up. In Section 2.3.1 I will discuss how LSCD methods make use of the different ways of meaning modeling, after which I will outline the different metrics to measure change in Section 2.3.2. Next, I describe which evaluation procedures have been used to evaluate the different methods (2.3.3).

2.3.1 Meaning Representation Techniques for LSCD

The different techniques used in LSCD methods to obtain functional word meaning representations can be directly translated to the distinction in language models discussed in Section 2.2. In other words, the leading existing LSCD methods differ from each other by using different language modeling techniques including count-based models, static word embedding models and contextualized language models. In addition, for each type of model there are different techniques that aim to ensure that the meaning representations of a word obtained from different subcorpora can be compared fairly with each other. Figure 2.3 shows a taxonomic overview of the different techniques used in the existing LSCD methods. I will now explain per type of language model which studies have made use of the different techniques in what way.

Using Count-based Word Representations for LSCD

One of the most prominent studies in the field of Lexical Semantic Change Detection is that of Hamilton et al. (2016b), who was one of the first to show how word vector representations derived from language modeling techniques can be used to reveal semantic evolution. Hamilton et al. (2016b) evaluated different types of word representations against historical changes in the form of statistical laws. One of the three types of word representations they used, and later also Dubossarsky et al. (2017, 2019), were sparse PPMI vectors. A separate V-by-V PPMI matrix is generated for each

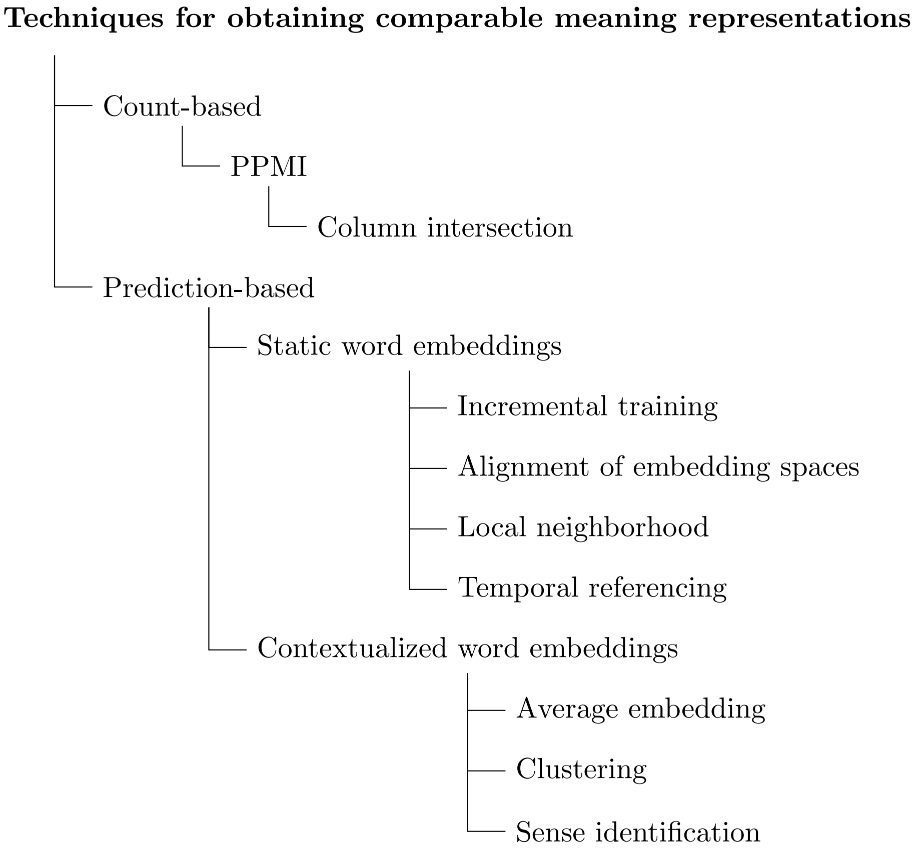


Figure 2.3: Taxonomic graph of different techniques for using word meaning representations in LSCD methods

time-specific subcorpus, where V is the size of the vocabulary of the subcorpus. Since the subcorpora are different, this generation results in matrices of different sizes. A word vector (a row) in one matrix cannot be directly compared with the same word in the other matrix. After all, the columns representing all context words do not match. To this end, the **column intersection** of the matrices in question is computed and subsequently the row vectors are limited to those columns.

Using Static Word Embeddings for LSCD

Using count-based models for semantic change detection is not as popular as using word embedding models. The central ideal of this type of methods is comparable to a count-based procedure and involves the construction of word embeddings for each corpus from a certain time period. In general, the algorithms used for the construction of word embeddings, like Word2Vec, involve random initialization of the (time-specific) semantic space, i.e. the trained embedding model. This means that semantic spaces developed by different training processes are not *aligned*. Even though a word has exactly the same meaning in different corpora, the numerical representations constructed by the models based on these corpora can be very different. Therefore, comparison of embeddings *within* a semantic space (intra-model comparison) is meaningful, but comparison of embeddings *between* different semantic spaces (inter-model comparison) is not. Different techniques have been introduced to overcome this problem.

One of the attempts to prevent the disalignment of embedding spaces was introduced by Kim et al. (2014), and also used by e.g. Noble et al. (2021), and boils down to an **incremental training** procedure of word embedding models. More specific, Kim et al. (2014) create embedding models that are initialized with the embeddings of the previous model, i.e. the embeddings trained on the corpus of the preceding time period.

The most successful kind of technique to date, with the work of Hamilton et al. (2016b) as prime example, is the **post-hoc alignment of embedding spaces** trained on different corpora. Hamilton et al. (2016b) use a technique called the orthogonal Procrustes, which is an optimization problem that takes two embedding models as matrices of word vectors and finds a matrix that maps the entries of one matrix most closely to the other while preserving the cosine similarities within each model.

Arguing that the alignment of semantic spaces is still vulnerable to distortions because perfect alignment is hardly feasible, methods have also been introduced that do not require alignment of semantic spaces at all, such as the methods by Wu et al. (2018) and Hamilton et al. (2016a). After training different word embedding models for the different time-specific corpora, Wu et al. (2018) construct a so-called topological semantic space in which each point (i.e. each word) has its own "semantic network". In this network, the nearest neighbors of the word, based on distance between the vectors in a time-specific embedding model, are positioned. Then, for a subsequent period of time, this space with neighbors is updated. The idea of Hamilton et al. (2016a) is conceptually similar. They create per target word a (distance based) **local neighborhood** for each temporal embedding model which they then compare.

A final type of method that is worth discussing, is the **Temporal Referencing** technique introduced by Dubossarsky et al. (2019), which also does not require alignment of semantic spaces. Instead of the separate training of embeddings for each time-specific corpus, all subcorpora are lumped together and all target words are replaced with variants indicating which time period the target word comes from. Subsequently, one semantic space can be created in which the target words can be distinguished and compared with regard to the different time periods.

Using Contextualized Word Embeddings for LSCD

A major disadvantage of using static word embeddings is the representation limitation of one meaning per word, which does not cover the polysemous nature of language. Changes in static word representations can reflect changes in the dominant meaning of the word but shifts in other meanings of the word remain undetected. In addition, it may also be the case that a change in vector representation is caused by a changing frequency distribution over the different senses, since static word embedding models are sensitive to word frequency (Noble et al., 2021). For example, with the growth of scientific literature, the meanings of, for example, the term 'culture' did not necessarily change, but the occurrence frequency did. The biological meaning of 'culture' (in the context of bacteria etc.) suddenly became much more frequent. The frequency relationship between this scientific significance and the social significance of 'culture' shifted considerably as a result. Static word embedding representations would reflect such a shift. Therefore, a static word embedding shift does not necessarily reflect a change in meaning. The recent development of powerful contextualized language models seems to be able to improve on these limitations, as the application of deep usage-based embeddings also being introduced into methods for the analysis of lexical semantic change.

Hu et al. (2019) were the first to construct sense representations with deep contextualized word embeddings to detect semantic shifts. In order to derive a representative number of senses from the large number of different usage representations of a word, they make use of the Oxford Dictionary. For each word they construct a representation for every dictionary sense by feeding pre-trained BERT a set of example sentences illustrating the sense in the dictionary and then extract the embedding representation from the last hidden layer of BERT. Subsequently, the sense of a certain usage-specific word representation in a target corpus is **identified** on the basis of cosine similarity. Then, the usage proportion of each sense in a given time period is calculated. In contrast to static word representations, this result does provide insight into shifts over different meanings of one word. Let me illustrate this with the previous example of the word 'culture'. The sense-based method would result in a usage distribution across the different meanings of 'culture' rather than a single word embedding for each time period. Before the rise of scientific literature, the usage share of the biological meaning relative to the social meaning of 'culture' would be much lower than in the subsequent time period. This result thus distinguishes between frequency shift and meaning shift.

Other techniques used in lexical semantic change detection methods to extract usage representations and induce word sense representations from deep contextualized language models were introduced later by Kutuzov and Giulianelli (2020), Giulianelli et al. (2020) and Martinc et al. (2020). Giulianelli et al. (2020) extract usage representations by summing the embeddings of all BERT's hidden layers dimension wise, using the pre-trained model. Subsequently, they induce different senses by applying K-means **clustering** to all usage representations of each word, where the parameter K is determined by an optimization technique that finds a value for K with high intra-cluster and low inter-cluster similarity. This clustering technique is applied over the complete corpus, i.e. including all time intervals, and then the probability distribution of usage representations over the clusters is determined for each time period. So this method, like that of Hu et al. (2019), also provides information about the use of different meanings of a target word in a certain period of time. Only now the different meanings are not predetermined but derived from the test data of all time periods.

In contrast to the previous methods, Martinc et al. (2020) fine-tune BERT on all time-specific corpora first to ensure domain adaptation. Subsequently, they extract usage representations by summing the last four encoder output layers in the fine-tuned BERT model for every word. Time-specific sense representations are then formed by taking the **average** of all usage representations from a certain time interval. Despite the expected benefits of the power of contextualized models compared to static word embedding models, the representation is still limited to one sense per token. However, so far this does not seem to be necessarily performance-degrading on LSCD tasks, which I will discuss further in the next section.

Earlier work by Giulianelli (Kutuzov and Giulianelli, 2020) experimented with both the cluster- and average-based applications of contextualized representations. Instead of summation of layers, the top layer, the average of the last four layers and the average of all hidden layers were also tried as contextualized representation. The experiments were performed not only with BERT but also with ELMo, and for each type of model, both the pre-trained variant and the variant fine-tuned on the union of test corpora were evaluated. In Section 2.3.3 I will further elaborate on the evaluations of different implementations of LSCD methods and their outcomes.

2.3.2 Measurement of Semantic Change

After constructing vector representations or probability distributions as semantic modeling of a particular word in different time-specific corpora, the task is to quantitatively determine the difference between them. Depending on how word meaning is modeled in the existing methods, different metrics have been used for this. Figure 2.4 shows an overview of these different metrics applied in the different approaches discussed in the previous section.

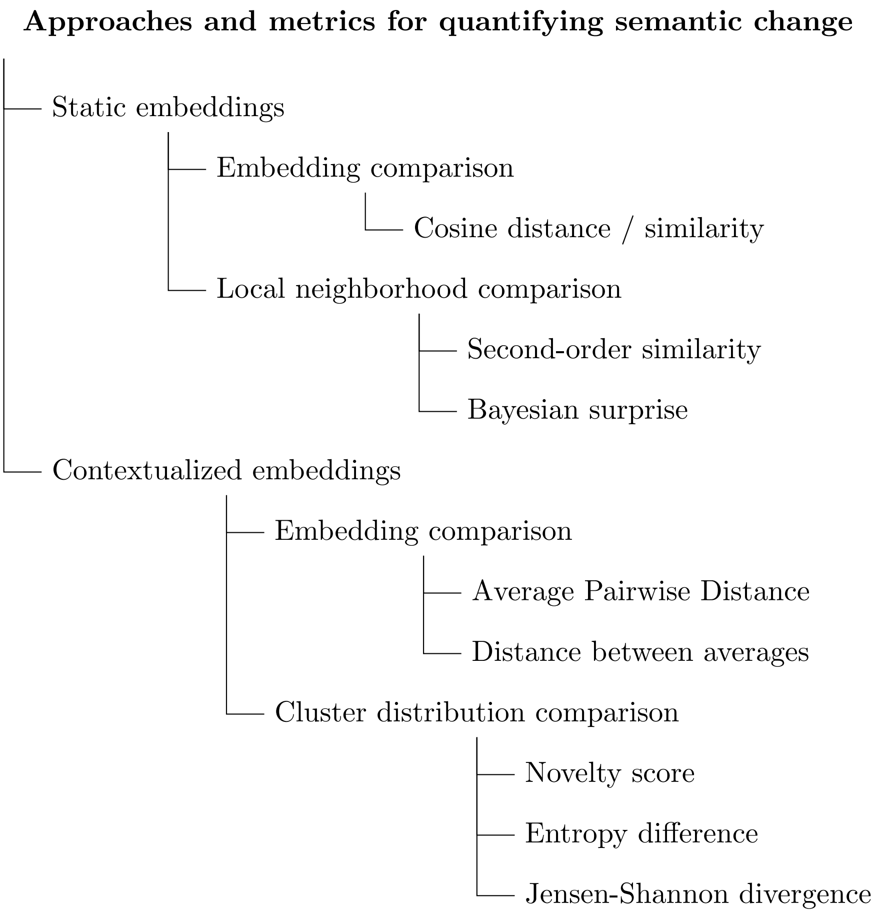

Approaches and metrics for quantifying semantic change
Static embeddings
Embedding comparison
Cosine distance / similarity
Local neighborhood comparison
Second-order similarity
Bayesian surprise
Contextualized embeddings
Embedding comparison
Average Pairwise Distance
Distance between averages
Cluster distribution comparison
Novelty score
Entropy difference
Jensen-Shannon divergence

Figure 2.4: Taxonomic graph of different approaches and metrics for quantifying semantic change in LSCD methods

Static Word Embeddings

As described earlier, for vector representations of word meaning, it is possible to calculate similarity using a geometric distance measurement such as the commonly used **cosine distance** (Kim et al., 2014; Hamilton et al., 2016a,b; Dubossarsky et al., 2017). A common derivation of this distance for use in NLP tasks is the **cosine similarity**. Given that the greater the distance between two word vectors, the smaller the semantic similarity is (and vice versa), the cosine similarity is defined by: $CosineSimilarity = 1 - CosineDistance$. A smaller cosine similarity, or a larger cosine distance, between vectors representing the same word in different time periods then identifies a shift in meaning. Dubossarsky et al. (2017) found that "naive cosine change is inherently biased towards words that appear in more variable contexts". This bias is associated with the previously mentioned word frequency effect, i.e. the cosine similarity between two word vectors tends to be lower for words with a lower frequency in the corpus. To overcome this bias, Noble et al. (2021) introduced the rectified change metric. Roughly speaking, this metric reflects the extent to which the measured cosine distance relates to the expected distance in a situation where no semantic shift has taken place at all. The rectified metric boils down to comparison with a control condition setup, as was introduced by Dubossarsky et al. (2017) as part of an evaluation step for LSCD methods, and will be discussed in the next section.

The methods in which the semantic representation of a word is not represented by a vector but by means of nearest neigbors, i.e. a local neighborhood, requires a different type of change quantification. Hamilton et al. (2016a) calculate the shift between the local neighborhoods of a word in different time periods using so-called **second order similarity**. They generate a second order vector for each local neighborhood that represents the cosine distances between the target word and each nearest neighbor from the union of local neighborhoods. Next, they examined to what extent the meaning of the target word shifts by calculating the cosine distance between the second order vectors (which reflects the extent to which the distances to the nearest neighbors for the target word change).

This quantitative measurement of semantic shifts provides information about the degree of shift, but not whether this is also statistically significant. Wu et al. (2018) introduce a metric that does cover this: the **Bayesian surprise** metric. Based on Bayes' theorem, probability functions are used to turn a time-specific local neighborhood into the *prior* distribution, and the local neighborhood in a subsequent time period into the *posterior* distribution. By means of the difference between these distributions (computed with the Kullback-Leibler divergence), the Bayesian surprise metric reflects the extent to which the posterior is a surprise with respect to the prior distribution.

Contextualized Word Embeddings

Measuring difference in distributions as in the Bayesian surprise metric is also applied in LSCD methods using contextualized word embeddings. As discussed earlier, one way to compare multiple contextual embedding representations of a target word, i.e. multiple word senses, in one subcorpus with the sense embeddings of the target word in another subcorpus is to evaluate the probability distributions across word senses. Hu et al. (2019) used the **Novelty score** for this. This metric, which was introduced by Cook et al. (2014), is defined by the division of the usage proportion of a particular sense in one subcorpus by the usage proportion of that sense in another subcorpus. Hu et al. (2019) take the maximum novelty score of all senses of a word as the change quantification value. With this approach, the results are sensitive to variation in discourse topic between the subcorpora, which, as discussed earlier, does not necessarily reflect a change in word meaning.

Giulianelli et al. (2020) use two metrics to measure the variation in relative prominence of senses of a particular word across time periods. The first, the **Entropy Difference**, calculates the difference in entropy between two time-specific word sense probability distributions. The entropy in this case reflects the uncertainty in sense interpretation of a word. The idea behind this is that the more polysemous the word is, the more uncertain the sense is. So if the interpretation of a word becomes more uncertain, the word would take on more different meanings, resulting in a high entropy difference, and vice versa. According to Giulianelli et al. (2020), a high entropy difference would therefore imply broadening of a word's meaning, and a negative value would imply narrowing. The **Jensen-Shannon divergence** (JSD) is the second metric, which was also used by Kutuzov and Giulianelli (2020). In the quantification of similarity between word sense probability distributions, this metric takes into account not only *that* proportions change, but also *which* sense proportions change. Thus, a high JSD value indicates that certain meaning clusters of a particular word have grown or shrunk, while a low value indicates that the size of the different meaning clusters, and thus the degree of use, has remained more the same over time.

Another way to measure change in a word's contextualized representations between different subcorpora is the **Average Pairwise Distance** (Kutuzov and Giulianelli, 2020; Giulianelli et al., 2020). This metric does not take different meanings of a word, or it's polysemy, into account but quantifies the mean shift in meaning across all contextualized representations of a word. For each possible pair between contextualized representations of a word in one and in the other subcorpus, a geometric distance is calculated, such as the cosine distance, and the average is taken. An alternative to this metric is to first compute the average word embedding for each subcorpus and then calculate the **distance between the average embeddings**. Both Martinc et al. (2020) and Kutuzov and Giulianelli (2020) applied this metric using the cosine distance (inverted or not).

2.3.3 Evaluation of LSCD Methods

Until recently, due to a lack of high-quality large-scale gold data, the evaluation of developed methods was one of the most difficult components within the field of Lexical Semantic Change Detection. Most studies evaluated their developed method(s) on a small scale, each in its own way, and against one or a few baseline methods (if there were any at all). I will discuss these approaches first, after which I will go into the recent first contributions to larger-scale gold data and method comparisons, and the results thereof.

Small-scale Evaluations

Many studies employed an evaluation procedure based on small-scale human assessments. These evaluations were often done in a **hypothesis-driven** way or (in combination with) a **data-driven** way. In the hypothesis-driven approach, an LSCD-method is applied to a small set of manually selected target words (one to several dozen) that are known to have undergone historical semantic shifts (Hamilton et al., 2016b; Wu et al., 2018; Martinc et al., 2020). Both Hu et al. (2019) and Giulianelli et al. (2020) used a list of one hundred English words annotated with degrees of semantic shift. This list was created by Gulordava and Baroni (2011) who asked human annotators to rank the words on a 4-point scale of semantic shift in a period of 40 years. The resulting list is therefore only specifically applicable to English-language data from the period ±1970 - 2010, if that is the case at all, given the fact that the ratings are not based on data but on intuition.

The second approach in small-scale evaluation is not based on a priori hypotheses. After applying an LSCD method to all words in the vocabulary, the top K words that underwent the greatest semantic shift according to the results are post-hoc verified by human assessments (Kim et al., 2014; Hamilton et al., 2016b). K is often a small number (e.g. 10) for reasons of feasibility.

Dubossarsky et al. (2017) argue that a detected semantic shift corresponding to human judgments does not necessarily indicate a valid method: "In order to establish the validity of an observation about meaning change, the result obtained in a genuine experimental condition should be demonstrated to be lacking (or at least significantly diminished) in a control condition.". The use of such a control set-up, as already referred to in Section 2.3.2, was also advised by Sommerauer and Fokkens (2019) who proposed guidelines for studying conceptual change using embedding models. Specifically they advised to use control words that should not exhibit conceptual change. Another control condition can be established with a similar corpus as the original one only without semantic shifting of the target word(s). Dubossarsky et al. (2017) artificially generate such a control condition by treating one subcorpus, which reflects the language of one time period so that it can be assumed that there are no semantic shifts in it, as the original diachronic corpus. This one subcorpus is thus divided into subcorpora and the LSCD methods are applied, which, in order to be judged as valid, should give a different result.

The artificial aspect of the latter approach is in line with a final direction of evaluation used in the absence of gold-annotated data, which is the use of **synthetic** data. The central idea behind this approach is to *simulate* a change in word meaning by manipulating data. Wu et al. (2018) do this, in addition to a small-scale hypothesis-driven evaluation, following a manipulation method introduced by Kulkarni et al. (2015). They take a word pair (with similar and sufficient frequency) and in one subcorpus of a selected time period they swap these words. Compared to a subcorpus of a contiguous time period (Wu et al., 2018) or an unmanipulated duplicate of the subcorpus (Kulkarni et al., 2015), the manipulated corpus shows a substantial change in meaning of these words, which a good LSCD method should detect.

Larger Scale Comparisons

The evaluation procedures for LSCD methods described above generally involve one or a few developed methods, compared to one or a few baseline methods. This makes it difficult to determine which methods in the field are better than others. Some studies contributed to a larger-scale systematic comparison of LSCD methods. Shoemark et al. (2019) compared methods based on static word embedding models and did so using a synthetic evaluation framework. For a sample of empirical data from Twitter, assuming no historical lexical semantic change, word replacements are applied in different ways to simulate different forms of semantic change. Schlechtweg et al. (2019) used two small-scale evaluation datasets that contained the human annotation of the semantic change of 22 German target words. They compared a slightly larger variety of methods than Shoemark et al. (2019), by also including different types of embedding models (besides SGNS also count-based models) and also the topic-based method of Frermann and Lapata (2016), in addition to several (non-)alignment techniques (such as orthogonal Procrustes and incremental initialization) and various change metrics (including cosine distance and Jensen-Shannon divergence).

One of the first contributions to large-scale high-quality human annotated evaluation data came with the first SemEval shared task on LSCD in 2020. Schlechtweg et al. (2020) presented an evaluation framework for LSCD with the first large-scale gold dataset for four languages: English, German, Latin and Swedish. For each language, 31 to 48 target words were assessed by 5 to 10 annotators in two corpora from two different time intervals. 100 usages of each target word from each time interval (i.e. 200 in total) were sampled and the semantic similarity between random pairs of these usages were rated on a four-point scale from unrelated to identical.

Both the results of the shared task (Schlechtweg et al., 2020) and the comparative studies of Schlechtweg et al. (2019) and Shoemark et al. (2019) show that methods that involve static word embedding models, trained with the SGNS algorithm, align semantic spaces with orthogonal Procustes, and measuring semantic change using the cosine distance, perform best on the LSCD task. It is striking that for this NLP task, methods based on static word embedding models outperform methods based on contextualized language models, while the latter proves to be much more powerful and substantially better performing for almost all other NLP tasks. Schlechtweg et al. (2020) give three possible explanations for this surprising result. The form of the test data could be an explanation as the examples were lemmatized and the context for each word included was fairly restricted while the contextualized language models are pre-trained on raw (not lemmatized) and context rich data. Secondly, contextualized language models and certainly the transformer-based models are new developments about which much is still unclear and needs to be explored, such as the fine-tuning possibilities and the selection and summary of hidden layers (for a single meaning representation). A final important and possible explanatory distinction between static and contextualized word embedding models is the fact that models like BERT are not solely trained on the focus corpus but are already pre-trained on a wide variety of data. This could mean that the word representations ultimately analyzed in the LSCD task contain more and perhaps irrelevant information that may mask the semantic shifting effect.

2.4 From Diachronic to Synchronic Variation

An underexposed aspect with regard to LSCD methods is the fact that they can be of value not only for the detection of diachronic semantic change, but also for synchronic semantic change, or rather variation. After all, change is variation over time. Del Tredici and Fernández (2017) and Lucy and Bamman (2021) are two studies, and to my knowledge the only two, that apply methods developed for Lexical Semantic Change Detection to detect semantic variation between different social communities. The principle behind this is to replace the time variable with socio-cultural based variables. To illustrate this, Del Tredici and Fernández (2017) use a method similar to the Temporal Referencing method of Dubossarsky et al. (2019), only replacing the temporal component with a community variable. This means that one Word2Vec model is trained on all community-specific subcorpora, in which the target words are labeled with the origin community. This results in one word embedding *per* community for each target word. The resulting word embeddings are all part of the same semantic space. Del Tredici and Fernández (2017) apply this method to data from various online communities on the social media platform Reddit. Lucy and Bamman (2021) use the same data source, and include many more communities. They analyze whether word use of a particular community deviates from the norm by comparing the sense representations of a word in a particular community with the representations in all communities. They use the state-of-the-art model BERT, and apply a cluster-based approach like Giulianelli et al. (2020) did in a diachronic setting, only using a PMI-based metric for quantifying the distinctiveness. They evaluate the results against community glossaries, which Reddit users create to define their ‘own language’ for newcomers to the community.

2.5 Conclusion

The literature discussed in this chapter provides the context for this thesis, which brings together different scientific disciplines. The new field of Cognitive Sociolinguistics emphasizes the importance of social factors for semantic variation. In accordance with the Distributional Hypothesis, the study of social variation usually involves empirical methods centered on the analysis of language use. Approaching semantics from a usage-based approach also gave rise to the emergence of language models within Computational Linguistics. These models evolved from simple count-based models to deep neural networks generating contextual word representations. Within Historical Linguistics it has already been shown how language models can be used for the analysis of semantic shifts through the development of Lexical Semantic Change Detection methods. A variety of introduced methods include different ways to generate word meaning representations and measure difference between them in a diachronic setting. Surprisingly, so far static word representations (e.g. Word2Vec-based) seem to perform best on LSCD tasks, and not the recent and generally successful deep contextualized models like BERT. Much is still unclear about the use of this last type of model for the analysis of semantic variation.

Both Del Tredici and Fernández (2017) and Lucy and Bamman (2021) show that LSCD-based methods can be used to gain new insights into semantic variation at community level. This thesis builds on the application of LSCD methods to social phenomena by further exploring the potential of using language models for Cognitive Sociolinguistic research. In addressing this issue, special attention is given to the need for greater understanding of state-of-the-art contextualized language models. Furthermore, it seems that the field of Lexical Semantic Change Detection is mainly focused on English-language data, with some studies focused on German or Swedish. In this research I will make use of Dutch data, which aims to have two advantages. Firstly, choosing Dutch is favorable for the assessment of empirical data, since Dutch is my native language. Second, to my knowledge this study will make a *first* contribution to research on computational methods for the analysis of semantic variation for Dutch.

Chapter 3

Methodology

In this thesis I focus on the use of language models for the analysis of semantic variation between different social communities. We have seen that there is already a lot of work on methods for analyzing diachronic semantic shifts, although much is still unclear about the use of the state-of-the-art transformer based models. The application of the Lexical Semantic Change Detection (LSCD) methods for the analysis of synchronic variation instead of diachronic variation is also not well-established yet, especially for Dutch data. I address these issues by selecting and applying two LSCD methods to the analysis of semantic variation between two Dutch-speaking social communities that differ fundamentally in political affiliation. To this end, I use data from Reddit. I evaluate the methods by combining different procedures applied in previous work on LSCD methods. I compose a small list of carefully selected target words for which concrete expertise-informed hypotheses exist regarding conceptual variation between the communities. Additionally, I implement different control set-ups to establish the validity of the results on these target words. In the following sections I will successively explain the data collection (Section 3.1), the formulation of the hypotheses (Section 3.2) and the experimental set-up to test these hypotheses with language modelling techniques (Section 3.3).[1]

3.1 Data

I compiled a dataset based on open source language material from Reddit, a social media platform made up of a large number of communities called subreddits. Within a subreddit people with shared interests can post and comment on specific topics. An effect that should ideally be excluded when analyzing semantic variation in relation to sociocultural variables is the appearance of difference in word meaning caused by discussing different topics. To limit this effect I selected two communities from the same domain, i.e. politics. Although from different convictions, the communities discuss similar topics as they both focus on political current affairs.

I scraped the comments from two Dutch subreddits corresponding to two social communities positioned at opposite ends of the Dutch political spectrum: `Poldersocialisme` and `Forum_Democratie`. The `Poldersocialisme` community describes itself as:

[1] The code and data used for this thesis can be found at `https://github.com/cltl-students/SanneHoeken_RMA_HLT_Thesis`

> "*Uit het donker naar het licht, poldersocialisten kent uw plicht! De grootste Nederlandse linkse subreddit!*"

This community consists of supporters of the left-wing political movement in the Netherlands, which includes progressive parties such as the SP and GroenLinks, which are generally in favor of a greater role for the government in social life. Members of Forum_Democratie define their subreddit as:

> "*Forum_Democratie is de onofficiële fan-made subreddit voor FvD-fans. Hier kunt u alles van meems tot artikelen of discussies gerelateerd aan FvD plaatsen. De subreddit is niet aan FvD zelf gelieerd.*"

This community includes supporters of one specific right-wing party, Forum voor Democratie (FvD). In line with right-wing ideology, this is a conservative party that wants to limit the role of government in social life. More specifically, FvD can be characterized as right-wing nationalistic with the main program points of introducing various forms of direct democracy such as referendums and elected officials, and strengthening national sovereignty.

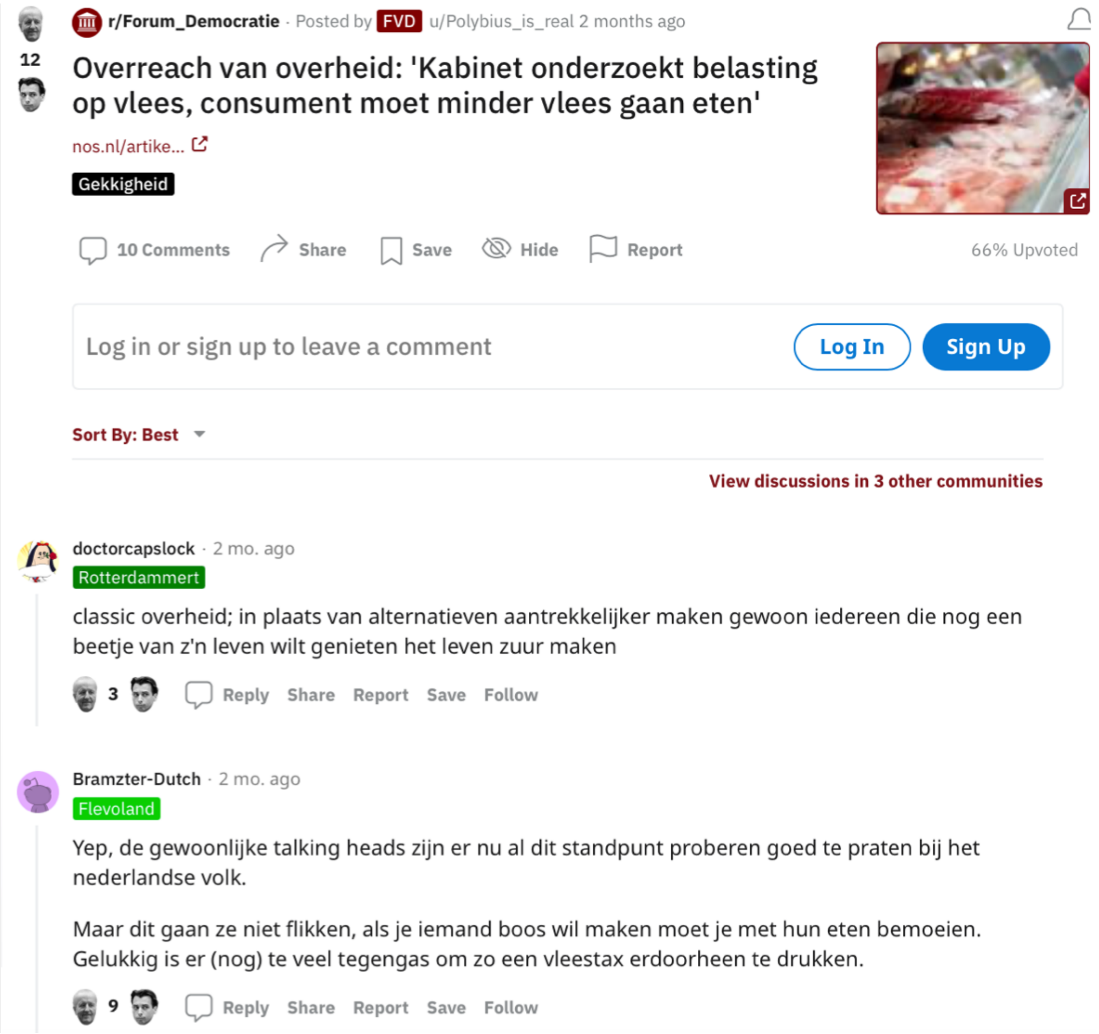
On Reddit, two main post types can be distinguished. Each member can post a *submission* in a subreddit which contains a title component and a content component. On every submission *comments* can be given (which consist only of a content component). Comments can again be commented, allowing a form of conversation. In Figure 3.1 you can see an example of a submission (top box) in the Forum_Democratie subreddit on which two comments (bottom two boxes) were given by other members.

To extract the data of the subreddits I used the Pushshift platform, on which real-time updated Reddit data is collected and made available to researchers via an API. I have scraped all existing submissions and comments from the two subreddits discussed up to April 3, 2022. For the generation of the final dataset I have removed all URLs, text between square brackets and some HTML escape characters. Instances that did not contain any text with alphabetic letters after the pre-processing steps (such as posts with only an emoticon, image or media link) were not included. Ultimately, the data extraction resulted in four datasets, consisting of 499 submissions and 58k comments from Poldersocialisme and 1207 submissions and 149k comments from Forum_Democratie. Table 3.1 shows more detailed statistics of each of the collections.

	Poldersocialisme		**Forum_Democratie**	
Members	7.3k		13.0k	
Creation date	May 30, 2018		Dec 16, 2017	
	Submissions	Comments	Submissions	Comments
N	499	58548	1207	149676
Tokens	57514	2337836	159940	6498946
Average length (in tokens)	115	40	132	43
Authors	297	4039	533	5205

Table 3.1: Statistics of the data extracted from the two subreddits Poldersocialisme and Forum_Democratie, collected at 3 April 2022

An initial analysis of the data showed that submissions generally contain little member-produced text. That is, the vast majority of submissions quote a news item



r/Forum_Democratie · Posted by FVD u/Polybius_is_real 2 months ago

Overreach van overheid: 'Kabinet onderzoekt belasting op vlees, consument moet minder vlees gaan eten'

nos.nl/artike...

Gekkigheid

doctorcapslock · 2 mo. ago

Rotterdammert

classic overheid; in plaats van alternatieven aantrekkelijker maken gewoon iedereen die nog een beetje van z'n leven wilt genieten het leven zuur maken

Bramzter-Dutch · 2 mo. ago

Flevoland

Yep, de gewoonlijke talking heads zijn er nu al dit standpunt proberen goed te praten bij het nederlandse volk.

Maar dit gaan ze niet flikken, als je iemand boos wil maken moet je met hun eten bemoeien. Gelukkig is er (nog) te veel tegengas om zo een vleestax erdoorheen te drukken.

Figure 3.1: Screenshot of an example of a Reddit submission in the Forum_Democratie subreddit with comments

or other media content in the title. The content of the submission then includes the link to the relevant source. This can also be seen in the example given in Figure 3.1. It is therefore more effective for the experiments in this study to only use the comments datasets.

3.2 Hypotheses

A common evaluation of computational methods for the analysis of semantic variation is hypothesis-driven using a pre-established list of target words. As Hamilton et al. (2016b) and others did in a diachronic setting, I also propose testable hypotheses about conceptualization of words based on human expertise. In this research, instead of hypotheses about historical shifts, I concentrate on knowledge about variation in social settings. The procedure followed and the hypotheses resulting from this are discussed in the following two sections (3.2.1 and 3.2.2 respectively).

3.2.1 Procedure

Using my own human expertise on language use, enriched with a theoretical expertise given my study background in language and communication, I prepared a first proposal of target words based on a manual analysis of the data. For the sake of reproducibility, I aimed to do this analysis as consistently as possible and to comprehensively describe it hereafter.

First, I took a random sample of about hundred submissions from each community and determined which topics are covered. Then I chose a few keywords that are relevant to these topics. With a simple rule-based script I extracted all comments in which these keywords are mentioned and I read every xth comment, where x depends on the number of results and is selected in such a way that I read an average of about fifty comments per keyword. If analysis shows that other frequently discussed topics and/or keywords occur in the data, I added these to the selection and carried out the described steps. Topics or keywords that yield few relevant results were deleted again.

Based on the samples of comments, I aimed to interpret how the members of the different communities conceptualize the keywords through their use. For each topic I stored a handful of sample comments to illustrate my hypotheses which are added to this thesis as Appendix A. These interpretations remained subjective and based on an individual assessment of a small selection of data. Therefore, in order to increase the validity of the hypotheses to be formulated, I also consulted a social scientist with relevant expertise. I discussed the results of the manual analysis with Mariken van der Velden, an Assistant Professor of Political Communication in the Department of Communication Science at the Vrije Universiteit Amsterdam. Van der Velden not only informed me about the plausibility of my preliminary hypotheses, but also enriched them with her knowledge about the political communities.

3.2.2 Resulting Framework

Through the described procedure I ended up with a selection of topics with one to several target words for each. I have formulated hypotheses regarding the conceptualization of each target word by the two different communities. I have presented an overview of these results in Table 3.2. A few explanations and illustrations will hopefully make it clear how these results should be interpreted. Referring to Lakoff's theory, the formulated hypotheses can be translated into differences in framing, i.e. which different mental structures the target word evokes. For the observed topics, a distinction can be made in the nature of semantic variation. For the first six topics, namely climate, vaccination, immigration, media, tax and government, there seem to be concrete differences in the frames that the target words evoke. For example, within the Forum_Democratie subreddit, the climate (and especially referring to its change) is framed as an unjustified hysteria while members of Poldersocialisme mainly emphasize that it is an urgent problem that needs to be solved. The following examples from comments in the two communities illustrate this:

> "*het probleem is echter dat het overwegend niet de weldenkende nederlanders zijn die deze klimaathysterie voorstaan*" (Forum_Democratie)
> "*in een tijd waarin we met zijn allen zouden moeten vechten om het klimaat te redden*" (Poldersocialisme)

Immigration, on the other hand, is explained by the right-wing Forum_Democratie as a problem that takes place on a massive scale and leads to growing crime. While

Topic	**Target words**	**Hypotheses**	
		Forum_Democratie	**Poldersocialisme**
Climate	*klimaat*	hysteria	problem
Covid-19 vaccination	*vaccineren* *vaccinatie*	issue of freedom/ invasion of liberty	protection of the vulnerable
Immigration	*immigratie* *immigrant* *vluchteling*	massive, problem, crime	responsibility, diversity, innocent/victim
Media	*media*	source of misinformation	source of information
Taxes	*belasting*	theft	more tax on capital
Government	*overheid*	passive, limited role	active, bigger role
Political movements	*links*	negative	positive
	rechts	selective positive, scapegoat	negative
	socialisme *socialist*	negative	positive
	liberalisme *liberaal*	selective positive	negative
	kapitalisme *kapitalist*	positive	negative, exploitation

Table 3.2: Target words and hypotheses

leftists frame this concept as something for which society must bear responsibility as immigrants fulfill a victim role in the situation. These conceptualizations can be derived from comments such as the following:

> "*echte problemen zoals illegale immigratie, toenemende criminaliteit en afnemende veiligheid*" (Forum_Democratie)
> "*omdat vluchtelingen niet voor niets vluchten en wij als rijk land zeker verantwoordelijkheid voor andere dragen*" (Poldersocialisme)

For the target words related to the last topic, political movements, the hypothesized differences appear to be of a less concrete nature, as they appear to differ mainly at the connotation level. In line with the affective polarization theory of Harteveld (2021) discussed in Chapter 1, members of Poldersocialisme have a positive interpretation of target words *links* and *socialisme* while the Forum_Democratie members dislike these concepts.On the other hand, for leftists, words like *rechts* and *liberalisme* carry negative connotations. Incidentally, the supporters of the Forum voor Democratie party do not have a convincingly positive interpretation of these words, as they also dislike other right-wing and liberal parties than themselves. They are also less positive about the right because they feel the right-wingers (like themselves) are being treated as scapegoats. A somewhat more concrete hypothesis exists for the target word *kapitalisme*, which seems to be structurally interpreted by the Poldersocialisme members as exploitation. The following examples from the datasets of the different communities illustrate the discussed expectations about the connotation of some words related to political movements:

> "*krijg je opeens een random domme zure linkse opmerking naar je toe geslingerd*" (Forum_Democratie)
> "*socialisme is in zo veel opzichten een juiste stap naar wat landen als nederland zo groot heeft gemaakt*" (Poldersocialisme)
> "*mensen die roepen dat rechts het probleem is*" (Forum_Democratie)
> "*kapitalisme synoniem voor bittere armoede, uitbuiting en oorlog*" (Poldersocialisme)

3.3 Experimental Set-up

Different datasets and a list of human-assessed target words with hypotheses regarding the semantic variation between the datasets are the most common ingredients for Lexical Semantic Change Detection. Since this research aims to gain a better understanding of the use of state-of-the-art transformer-based language models for this task, I have selected an LSCD method of this type which I will describe in Section 3.3.2. Before that, in Section 3.3.1, I will discuss the selected baseline method that relies on a count-based language model. The different types of language models are used to derive word meaning representations from the different datasets. I will outline the metrics that I use to measure the difference between the representations in Section 3.3.3, after which I describe the different control setups to establish the validity of the measured results in Section 3.3.4.

3.3.1 Using a Count-based Language Model

As Schlechtweg et al. (2020) did to compare LSCD methods in the SemEval 2020 Task, this study also uses a count-based model as a baseline. To recap from Section 2.2.1, a common count-based language model applied in previous work is the PPMI model (Hamilton et al., 2016b; Dubossarsky et al., 2017, 2019), which I also chose to implement in this research. In contrast with predictive-based language models, which do result in more powerful embeddings, the sparse count-based vectors have the favorable property of being transparent. From these "explicit meaning representations" (Dubossarsky et al., 2017) it can be directly deduced on the basis of which observations words are more or less similar. After all, every value in the vector of a word corresponds directly to a context word. State-of-the-art language models, such as BERTje (de Vries et al., 2019), consist of deep neural network architectures with millions of parameters. This complexity makes explaining word meaning representations, and thus the justification for detected semantic variation, hardly feasible. So, although representing less rich information, PPMI vectors may provide more explanatory information than neural network alternatives.

In this study's experiments, the same procedure for a PPMI implementation is followed as in these previous works, only then applied in a social rather than temporal setting, and with some different preprocessing due to language. All comments in the datasets are lemmatized with the Dutch NLP pipeline from the open source library Spacy. All non-alphanumeric characters have been removed and the text has been lowered. Given the large number of compound words in the Dutch language, all words in the data in which the target word occurred were split, in order to cover as many usages of the target words as possible. Words like ‘klimaatverandering’ or ‘overheidsbemoeienis’ were thus split into ‘klimaat’ and ‘verandering’ and ‘overheid’ and ‘sbemoeienis’ respectively.

For each community-specific dataset d, i.e. the comments dataset of `Poldersocialisme` and of `Forum_Democratie`, with vocabulary size V_d, a V_d-by-V_d co-occurence matrix is computed. The window size was set to 10. Each of the co-occurence values is then converted to PPMI scores. To recall from the previous chapter, Pointwise Mutual Information (PMI) between a word w and word c measures how often w occurs in a context with c in a corpus, compared to co-occurence of the words by chance. In order to mitigate biases towards infrequent context words, in the computation of the PPMI scores, the context probabilities are smoothed with a parameter α by raising the probability to the power of α (Jurafksy and Martin, 2021). In my experimental set-up the value of α is set to 0.75 as this value leads to improved results according to Levy et al. (2015). The PPMI score, with α-smoothing, for a word w and context word c is defined as:

$$PPMI_\alpha(w,c) = \max\left(\log\left(\frac{P(w,c)}{P(w)P_\alpha(c)}\right), 0\right) \tag{3.1}$$

where $P(w,c)$ denotes the probability that word c appears as a context word of w and $P(w)$ and $P_\alpha(c)$ denote the marginal probabilities of the word and its context respectively. The latter, to which the alpha smoothing applies is then defined as:

$$P_\alpha(c) = \frac{count(c)^\alpha}{\sum_c count(c)^\alpha} \tag{3.2}$$

To align the resulting community-specific PPMI matrices, the column intersection of the matrices is computed and subsequently the row vectors are limited to those columns. The rows corresponding to target words are selected as meaning representations and compared between the different communities, the details of which are discussed in later sections.

3.3.2 Using a Contextualized Language Model

One of the main components of this research is the application of contextualized language models for the analysis of semantic variation. In the few previous studies (e.g. Giulianelli et al. (2020); Hu et al. (2019)) using pre-trained transformer-based language models in LSCD methods, BERT was the common choice. In the experimental set-up of this study, BERTje (de Vries et al., 2019), the Dutch variant of BERT, is used.

Before the comments in the different datasets were fed to BERTje, either for fine-tuning or embedding extraction, all words in the data containing the target word were split, just like in the PPMI approach. This seems superfluous at first since BERTje’s tokenizer tokenizes input text into subwords. Nevertheless, an analysis of the default tokenization results showed that compounds containing a target word are not always split in the most optimal way for the coverage of target words in the experiments. For example, the target word *overheid* still contains the connection morpheme ‘-s-’ after the default tokenization of the compound ‘overheidssteun’ by BERTje:

‘[CLS]’, ‘overheids’, ‘##steun’, ‘[SEP]’

A tokenization result like this would hinder the capture of the target word mention. Splitting the compounds beforehand solves this problem:

‘[CLS]’, ‘overheid’, ‘s’, ‘##steun’, ‘[SEP]’

As discussed in Chapter 2, Schlechtweg et al. (2020) indicated that pre-trained language models in the SemEval task may have suffered from lemmatized test data. Therefore, I experiment with lemmatizing the data before feeding it to BERTje, in addition to the main setting with raw data.

Following Kuzutov and Giulianelli (2020), BERTje's tokenization function is set in such a way that target words are never split into subwords and target words that are not in BERTje's vocabulary are added to it. The latter implies that BERTje has no pre-training information about these added words, unlike target words that were already in BERTje's vocabulary. An effect of this distinction, which is not addressed in Kutuzov and Giulianelli (2020), seems to be evident in the results of the experiments in this thesis. For the target words that had to be added to BERTje's vocabulary the measured similarity is found to be consistently lower than for pre-existing target words. That is why I decided afterwards to draw conclusions based only on the target words that were already in BERTje's vocabulary. This will be further explained in Chapter 4.

After the preprocessing and tokenizer configuration, the pre-trained BERTje model is fine-tuned to the union of community-specific datasets with the training objective of Masked Language Modeling (MLM) following the implementation of Kutuzov and Giulianelli (2020). Fine-tuning with this objective is algorithmically equivalent to pre-training BERTje. During pre-training BERTje learns to ‘understand’ general language by being trained on a wide variety of web data. Fine-tuning then provides domain adaptation, making BERTje more able to ‘understand’ language specific to the selected Reddit communities.

The pre-trained BERTje model is made publicly available through Hugging Face's transformers library for Python (Wolf et al., 2020). The maximum sequence length that BERTje can process is 512 tokens. During fine-tuning, input instances longer than 512 tokens are truncated, i.e. cut to the maximum length, and instances shorter were padded, i.e. supplemented with [PAD] tokens. The model was fine-tuned for 3 epochs with a batch size of 8 instances. All other training hyperparameters were set to the default values as Kutuzov and Giulianelli (2020) also implemented.

The usage representations of all target words are extracted by feeding fine-tuned BERTje each mention of each target word in it's context, which concerns the entire comment up to 512 tokens preprocessed as described above. 1693 comments in the datasets exceeded the length of 512 tokens. These were cut in such a way that the target word is always in the comment. The usage representation of each target word includes the concatenation of all hidden layers of the sentence position of the target word in the context. This resulted for each target word in a community-specific matrix in which each row is a usage representation of the target word in one of the datasets. The number of rows corresponds to the number of mentions of the target word and the number of columns to the dimensionality of the concatenated hidden layers.

Experiments with different set-ups (Kutuzov and Giulianelli, 2020) showed that fine-tuning the model as described proved to be the best performing strategy compared to pre-trained models only. Testing different selections and summary techniques of the hidden layers to derive a single word embedding seemed to make little difference to the detection task. I aimed to verify both outcomes by 1) running the experiments with the pre-trained version of BERTje as well, and 2) extracting not only the concatenation of all hidden layers, but also only the top layer as word embedding and doing the same experiments on this. Both variables showed no effect on the degree of variation detected between the different datasets. This is explained in more detail in Chapter 4.

3.3.3 Measuring Similarity

Given the PPMI representation of a target word in one community and the representation of the same word in the other community, the similarity in word meanings was computed using the Cosine Similarity metric. Between two word vectors x_i and x_j the Cosine Similarity is defined as:

$$CosSim(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|_2 \|x_j\|_2} \tag{3.3}$$

where $\| * \|_2$ is the 2-norm of its argument $*$, and $x_i \cdot x_j$ is the dot product of x_i and x_j.

For the contextualized word representations, which involve multiple representations for a target word per community, previous work (Giulianelli et al., 2020; Kutuzov and Giulianelli, 2020) shows that a summary embedding comparison leads to better results than measuring distance between cluster distributions (via e.g. the Jensen- Shannon divergence). In this experimental setup, the Average Pairwise Distance is used. For each pair of the two sets of usage representations of a target word w, i.e. matrix U_w with N usage representations x for each community, the distance d is calculated and the average is taken over all these distances:

$$APD\left(U_w^1, U_w^2\right) = \frac{1}{N^1 \cdot N^2} \sum_{x_i \in U_w^1, x_j \in U_w^2} d\left(x_i, x_j\right) \tag{3.4}$$

For d I take the Cosine Similarity as defined above. The Average Pairwise Distance based on the Cosine Similarity can be better called the Average Pairwise Similarity to avoid confusion with the Cosine Distance as the basis. I will therefore use the term Average Pairwise Similarity, abbreviated as APS, for this metric from now on.

To verify the effectiveness of the similarity metric, I also applied the metric to different target words in different communities. A valuable metric should show that the semantic similarity between a target word, e.g. *klimaat*, in one community and a semantically hardly related other target word, e.g. *liberaal*, in the other community should be much smaller than between the same target word in different communities.

3.3.4 Control Set-ups

Inspired by the useful control conditions and the data manipulation strategies for the evaluation of LSCD methods formulated by Dubossarsky et al. (2017) and Kulkarni et al. (2015) respectively (as discussed in Section 2.3.3), I included a combination of these strategies in the experimental set-up, which I will present in subsequent sections.

Within versus Between community

The main control set-up to establish the validity of the results on the target words boils down to a within-community versus between-community comparison. I have schematically shown the setup of this experiment in Figure 3.2. The dataset of one of the communities was randomly split into two datasets, resulting in three test datasets. By splitting the dataset with comments from the `Forum_Democratie` subreddit, the sizes of the resulting datasets are also more equal compared to the original ratio between the dataset sizes. Table 3.3 displays the new statistics of the datasets, including the

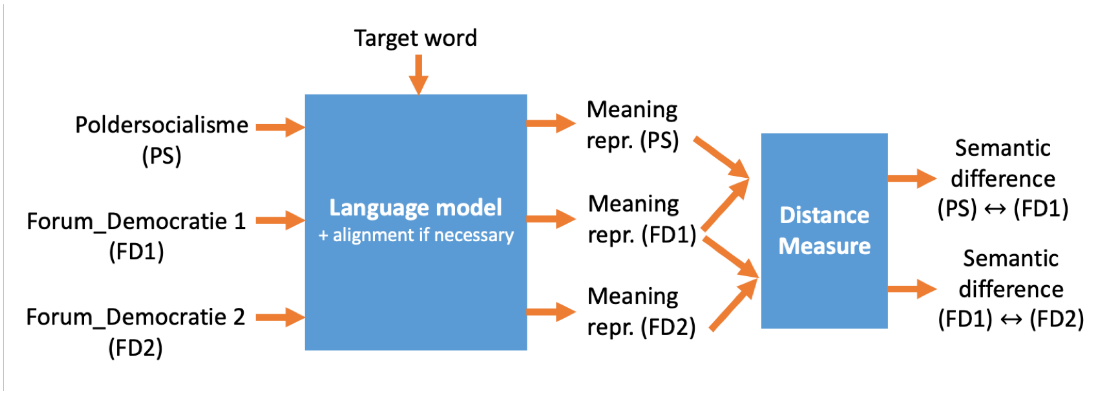


Figure 3.2: Control set-up for a within-community versus between-community comparison

number of mentions of each target word in each dataset after the compound splitting was applied to the raw (so non-lemmatized) data.

A between-community comparison is performed by applying the described methods to one of the Forum_Democratie datasets and the Poldersocialisme dataset. This can be considered as the original set-up, where it is tested whether the hypotheses about the semantic interpretation of the target words are detected by language model based methods. To verify that the detected differences in the between-community setting are an effect of difference in community and not some "noise", a similar experimental setting was created but with the community variable kept constant. The so-called within-community comparison measures difference between meaning representations between datasets of the same community, i.e. Forum_Democratie.

Control Words

If a method detected semantic variation which appears to be an effect of community difference, the question remains whether the observed difference in semantic similarity is specific for the hypothesized target words, or applies to more words, or even the general discourse, and thus to almost all words. To test this issue, the method has also been applied to a small set of control words. The control words should be words for which no semantic variation is expected between the different political communities. To this end, I selected fairly 'neutral' words for which I do not expect a substantial difference in semantic conceptualization based on the communities' difference in political preference. This is in contrast to the target words, which are related to current political topics about which the two political communities have clearly opposing views.

The selected target words with the number of mentions in the different datasets are shown in Table 3.4. I selected twelve words from different parts of speech (nouns, adjectives and verbs) that I believe occur in fairly everyday discourse and cannot be directly associated with a specific domain. The occurrence frequency of the control words appears to be in a slightly lower but comparable range as the target words. The number of mentions of the control words in the datasets is between 55 and 1990 and for the target words this is between 83 and 5546.

Lakoff's theoretical framework as introduced in Chapter 1, arguing the association between different worldviews and different word meanings, is actually applied here "in the opposite direction". That is, similar rather than different worldviews would

		Polder socialisme	**Forum_ Democratie 1**	**Forum_ Democratie 2**
	Comments	58548	74838	74838
	Tokens	2337836	3253695	3255251
Topic	**Target words**	**Mentions of target words**		
Climate	*klimaat*	1154	2839	2944
Covid-19 vaccination	*vaccineren*	83	893	797
	vaccinatie	99	729	684
Immigration	*immigratie*	204	1009	964
	immigrant	140	381	365
	vluchteling	312	499	468
Media	*media*	1042	2568	2582
Taxes	*belasting*	1282	1124	1254
Government	*overheid*	1531	2131	2107
Political movements	*links*	5546	5025	5026
	rechts	3106	3859	3894
	socialisme	1779	332	274
	socialist	3081	416	387
	liberalisme	358	154	149
	liberaal	1626	675	666
	kapitalisme	1890	239	271
	kapitalist	1662	146	148

Table 3.3: Statistics of test datasets including mentions of the target words

be accompanied by similar frames contributing to this, which would be reflected in similar language. In other words, words about which language users do not have different views do not allow to have contestable meanings. The selected control words are intended to satisfy this assumption, so it is expected that the measured semantic similarity between the control words in the datasets of different communities (between-communities) is 1) approximately equal to the semantic similarity between the control words in datasets of the *same community* (within-community), and 2) higher than the semantic similarity between the *target words* in datasets from different communities (between-communities).

However, these expectations may not come true according to the consultation of political communication scientist Mariken van der Velden. She expected the two political communities to talk differently about almost everything and couldn't really come up with control words for this reason. Contrary to the Lakoff-based assumption, the results of the experiments with the control words do indeed agree with Van der Velden's disclaimer. I will discuss this in more detail in the following chapters.

Data Manipulation

The last evaluation strategy in the experimental set-up concerns the simulation of semantic variation through data manipulation (following Kulkarni et al. (2015)) and

Control words	Polder socialisme	Forum_ Democratie 1	Forum_ Democratie 2
boek	763	686	588
stad	371	394	367
stemmen	1216	2071	1990
spreken	766	1424	1394
snel	749	1280	1379
nieuw	1279	1848	1809
lopen	661	1129	1133
tafel	95	185	194
blauw	55	92	109
slapen	83	65	67
muur	82	124	121
warm	61	147	141

Table 3.4: Number of mentions of control words in the test datasets

serves a dual purpose. First, just like the small-scale hypothesis-driven evaluation, this experiment should test whether the method is indeed able to pick up semantic variation. And secondly, this experiment attempts to address the influence of pre-training of the contextualized language models on the semantic variation detection results. To my knowledge this has not been done in previous studies.

As suggested by (Schlechtweg et al., 2020), the lower performance of pre-trained contextualized language models on LSCD compared to static word embedding models could be explained by the fact that the extracted word representations contain additional and masking information due to pretraining. For a target word in the context of a test instance, this would mean that the token embedding also represents a lot of information about the token in the large amounts of 'general' pre-training data. In fact, the extent to which pre-training information about the target word would be represented seems to overshadow information from the test data, which might reflect semantic variation. The question is therefore whether the context of target words in test examples offers enough predominance in the final representation of a word, or whether the pre-training information is more dominant. Through a data manipulation strategy I tried to test this.

In one of the datasets from the `Forum_Democratie` subreddit, all instances of a target word, the **donor** word, were replaced with all instances of another (barely semantically related) target word, the **recipient** word. The original donor instances in the dataset were replaced with a nonexistent/meaningless word. As a result, the instances of the donor word are in unnatural contexts. Nothing is changed in the other dataset, with comments from the `Poldersocialisme` subreddit. Table 3.5 shows data snippets to illustrate the manipulation results in the `Forum_Democratie` dataset. The recipient word *klimaat* has been replaced by the donor word *liberaal*, which has previously been replaced by the semantically empty placeholder *foo*.

If the context of the test instance *does* influence the word representations extracted from BERTje, the consequence of the described manipulation should be that the semantic similarity between the donor target word in the different communities should be substantially lower than in the original setting. In fact, if the context *is* the determin-

Original	**Manipulated**
dit is gewoon een advertentie van de klimaatgekkies	*dit is gewoon een advertentie van de liberaalgekkies*
roekeloos klimaatbeleid is wel degelijk gevaarlijk	*roekeloos liberaalbeleid is wel degelijk gevaarlijk*
de hysterische, maar vaak inhoudelijk vrij lege, weinig concrete klimaatgekte, die nu de mainstream lijkt te worden	*de hysterische, maar vaak inhoudelijk vrij lege, weinig concrete liberaalgekte, die nu de mainstream lijkt te worden*
het is zuiver liberaal dus dat verdient een klein bloemetje	*het is zuiver foo dus dat verdient een klein bloemetje*
ik ervaar het FvD als een laatste kans voor een vrij, liberaal en democratisch Nederland	*ik ervaar het FvD als een laatste kans voor een vrij, foo en democratisch Nederland*
een liberaal wil de collectieve sector zo klein mogelijk en gescheiden van de economie	*een foo wil de collectieve sector zo klein mogelijk en gescheiden van de economie*

Table 3.5: Snippets from the original and manipulated version of the `Forum_Democratie` dataset

ing factor, it would be expected that the measured similarity would move more towards the value of the result in the original (unmanipulated) setting between the recipient word in the `Forum_Democratie` dataset and the donor word in the `Poldersocialisme` dataset.

The extent to which word embeddings of contextualized language models are sensitive to context of test data instances has been addressed before by Ethayarajh (2019) with the question: "How Contextual are Contextualized Word Representations?". One of the key findings of this work relevant to this thesis concerned the context-specificity experiments, which showed that the upper layers of (among others) BERT are more context-specific than the lower layers. Therefore, the data manipulation experiment is also performed with extracting only the top layer as word embedding, instead of concatenation of all hidden layers. In line with the findings of Ethayarajh (2019), the expectation is that in the manipulated setting the similarity between the community-specific representations formed by only the top layer (which would thus be more sensitive to context) is smaller than between the representations formed by the concatenation of all hidden layers.

Chapter 4

Results

In this chapter the results of the experiments as described in Chapter 3 are presented and discussed. These experiments serve the purpose of this thesis to investigate the possibilities of language models for the analysis of semantic variation from a socio-cognitive perspective. First of all, the results of the baseline Lexical Semantic Change Detection method based on a PPMI language model in combination with the main control set-up are presented (Section 4.1). This set-up includes the comparison of the between-community and within-community results, performed on the pre-established list of target words. Next, the same type of results are discussed for the method based on contextualized representations extracted from BERTje (Section 4.2). Sanity checks into the effectiveness of the chosen metrics to quantify semantic difference are then briefly reported (Section 4.3). The results of the other two control experiments are described in Section 4.4. First, further validation of the detected variation is addressed with the results on the control experiment using control words (Section 4.4.1). The results of the final control experiments involving data manipulation, aimed at a better understanding of the functioning of a method based on a pre-trained contextualized language model, are presented in turn (Section 4.4.2). Finally, I return to the results of the count-based language model as it can, due to its transparency property, provide more refined information about the semantic variation through nearest neighbor analysis and representation values (Section 4.5).

Dataset	Abbreviation
Poldersocialisme	PS
Forum_Democratie 1	FD1
Forum_Democratie 2	FD2
manipulated	mnp
Cosine Similarity	CosSim
Average Pairwise Similarity	APS

Table 4.1: Abbreviations used for the reported results in this chapter

Before presenting the results, I make a few reading guide notes that that hold for all of the following reported results. The first concerns the between-community vs within-community experiments, in which results on datasets from different communities are compared with datasets from the same community. The between-community comparison reported in this chapter consists of the comparison of word representations from the Poldersocialisme dataset and the Forum_Democratie 1 dataset. All between-community experiments were also done for the Forum_Democratie 2 dataset. These results showed that all results were consistently similar when comparing Poldersocialisme representations with those of Forum_Democratie 1. For repetition reasons, these results have therefore been omitted from the main report.

Furthermore, for reasons of space, abbreviations for the datasets corresponding to the different communities and other terms that apply to the remaining chapter are used to report the results in the tables, which are defined here once in Table 4.1 below.

4.1 Using a Count-based Language model

After creating PPMI matrices for each of the community-specific datasets, the matrices were aligned through column intersection. As explained in the previous chapters, PPMI vectors are transparent in that each value in the vector corresponds to a word in the vocabulary of the training dataset. For a fair comparison of vectors derived from different datasets with different vocabularies, the resulting vectors should be aligned. Column intersection solves the problem that words not existing in both datasets cause a mismatch of the vectors' dimensions.

Target words	CosSim	
	Between	Within
klimaat	0,14	0,31
vaccinatie	0,11	0,22
immigratie	0,09	0,18
vluchteling	0,18	0,28
media	0,11	0,20
belasting	0,24	0,33
overheid	0,17	0,27
links	0,10	0,20
rechts	0,11	0,20
socialist	0,08	0,20
liberaal	0,06	0,21
kapitalisme	0,08	0,23
Average	0,12	0,24

Table 4.2: Cosine similarity between PPMI vectors of target words in different datasets, with the Between-column comparing datasets from different communities, and the Within-column comparing datasets from the same community.

Next, the PPMI representations of each target word formed by a row in the matrix were extracted. The similarity between the representations of the same target word for the different communities was then measured with the cosine similarity metric. Table 4.2 shows the results of the cosine similarity measures between the PPMI vectors of the same target word in different communities. It can be seen that the cosine similarity between PPMI vectors in datasets from the same community is consistently higher for all target words than between the vectors in datasets from different communities. This is in line with expectations, since 1) this implies that there is a relatively low semantic similarity between the representations in datasets corresponding to different communities, and 2) this effect could be attributed to the community variable since the effect is substantially less when comparing datasets from the same community.

Although the observed pattern clearly applies to all target words, there is still some variation between the different target words. For example, according to these results, the difference in semantic similarity between the within-community and between-community comparison is somewhat smaller for the target words *immigratie*, *media*, *belasting* and *rechts*. A larger difference exists for the target words *klimaat*, *liberaal* en *kapitalisme*. A greater difference in semantic similarity would thus indicate a greater difference in conceptualization of those target words between the members of the `Forum_Democratie` and `Poldersocialisme` subreddits, and vice versa. More fine-grained information on the measured variation follows in Section 4.5.

4.2 Using a Contextualized Language Model

So far, the baseline method seems to confirm the formulated hypotheses. It is now interesting to see how the method based on a more advanced language model, which can compute richer and more powerful representations of word meaning, compares to this. The Dutch BERT-based model, BERTje, was used to extract contextualized representations for each mention of the target word in the different datasets. The reported results in this section concern the results for the representations, formed by the concatenation of all hidden layers, extracted from BERTje fine-tuned to the union of all communities' raw datasets, so preprocessed as described in Chapter 3 without lemmatization. Table 4.3 shows the results of the Average Pairwise Similarity measures between the token embeddings of the same target word in different datasets.

Target words	APS	
	Between	Within
klimaat	0,78	0,78
vaccinatie	0,78	0,79
immigratie	0,79	0,79
vluchteling	0,80	0,80
media	0,77	0,77
belasting	0,76	0,76
overheid	0,78	0,78
links	0,77	0,77
rechts	0,75	0,75
socialist	0,78	0,78
liberaal	0,76	0,76
kapitalisme	0,79	0,78
Average	0,78	0,78

Table 4.3: Average Pairwise Similarity between token representations of target words in different datasets, with the Between-column comparing datasets from different communities, and the Within-column comparing datasets from the same community.

In contrast to the results of the experiments using a PPMI model, there is no difference between the within-community and between-communities results for any of the target words. Thus, the embeddings derived from BERTje do not seem to reflect the community-specific semantic variation as hypothesized. This becomes even more visible from a visualization of the word embeddings. Figure 4.1 shows an example of this. All token embeddings of the target word *vluchteling* are projected on a two-dimensional map using the statistical method t-SNE (van der Maaten and Hinton, 2008).

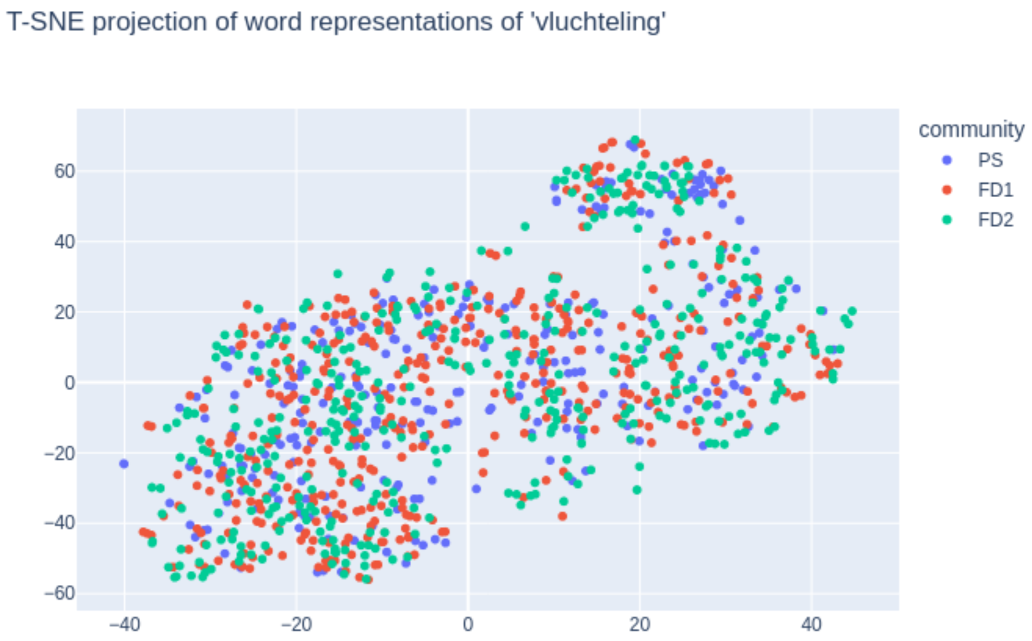


Figure 4.1: T-SNE projection of word representations of the target word *vluchteling*

A visualization of the token representations in line with the expectations would involve a large overlap with representations from FD1 (red) and FD2 (green), i.e. from the same community, but absolutely not with PS (blue), since the expectation is that the embeddings of the latter consist of other values reflecting other meaning elements, relative to the FD datasets. However, the actual visualization shows a large overlap of embeddings from all three datasets, with no difference in embeddings between any of the datasets compared to the other. Based on these results, no community-dependent semantic variation appears to exist.

Pre-trained or Fine-tuned

As mentioned in Chapter 3, experiments by Kutuzov and Giulianelli (2020) showed that the best performing strategy for contextualized language models deployed for Lexical Semantic Change Detection involves fine-tuning the union of all communities' datasets. I initially followed these recommendations and then tried to verify this myself by also performing the experiments with pre-trained BERTje.

The results show that the measured similarity between the pre-trained representations was generally lower than between the fine-tuned representations. A logical explanation for the overall higher similarity between fine-tuned representations could be that the model included more specific, and thus less broader, information in the hidden layer representations while the model was being fine-tuned, thus in fact undergoing domain adaptation. However, the pattern for the within-community versus between-community comparisons was exactly the same for the pre-trained model compared to the fine-tuned model. In other words, there was also no community effect for the semantic similarity between the pre-trained contextualized representations of the target words.

Table 4.4 summarizes the results for the within-community and between-community comparison of representations extracted from pre-trained BERTje and BERTje fine-tuned on the union of all datasets. More detailed results are included in Appendix B.

BERTje	Mean APS	
	Between	Within
Fine-tuned	0,78	0,78
Pre-trained	0,71	0,71

(*the Standard Deviation for each reported mean is 0,01*)

Table 4.4: Mean APS for representations extracted from pre-trained BERTje and BERTje fine-tuned on the union of all datasets

Pre-existence in BERTje's Vocabulary

The target words *vaccineren*, *immigrant*, *socialisme*, *liberalisme* and *kapitalist*, which were in the original target word list (presented in Section 3.2.2), did not exist in BERTje's pre-trained vocabulary. Following the method of Kutuzov and Giulianelli (2020) these words were added to BERTje's vocabulary before the model was fine-tuned and the embeddings were extracted. However, the results seem to indicate that this strategy is not without implication. Table 4.5 shows the average similarity measurements for the target words that already existed in BERTje's vocabulary (True) and for the target words that had to be added to BERTje's vocabulary (False). For each of the last set of words, i.e. the five target words listed above, the APS was calculated between the representations in the datasets of the same (Within) and different (Between) communities. This was done for the representations generated by the pre-trained as well as the fine-tuned version of BERTje. Then the average of these five APS measurements in each setting is taken, and these are the values reported in the False column. The True column shows the same kind of results, but for the other 12 target words that already existed in BERTje's vocabulary. More detailed results for each of the target words are included in Appendix B.

The Average Pairwise Similarity is found to be consistently lower for the words that were added to BERTje's vocabulary. Thus, the semantic similarity between word representations that already existed in BERTje's pre-trained vocabulary is relatively higher. A possible explanation is that these embeddings probably cover more common information, i.e. pre-training data, in their representations. The possible explanation for the higher similarity for words that exist in BERTje's pre-trained vocabulary is motivated even more by the results for the representations extracted from pre-trained BERTje. If no usage information about words is obtained in the fine-tuning phase, and thus the representations would be solely based on the sentence context and the pre-training information that is missing for words that do not pre-exist in BERTje's vocabulary,

Target words in Bertje's vocabulary	Mean APS			
	Fine-tuned		Pre-trained	
	Between	Within	Between	Within
True	0,78 (0,01)	0,78 (0,01)	0,71 (0,01)	0,71 (0,01)
False	0,74 (0,01)	0,74 (0,01)	0,60 (0,03)	0,60 (0,03)

(*Standard Deviation reported between brackets*)

Table 4.5: Mean APS between representations extracted from fine-tuned and pre-tained BERTje for target words that were already in BERTje's vocabulary (True) and target words that were added to BERTje's vocabulary (False)

the similarity drop is even more extreme. This is an interesting result to take in consideration in future design of semantic variation methods using pre-trained language models. However, as the results on the target words added to BERTje's vocabulary are likely to be less reliable, I have not included these words in reporting all previous and forthcoming experiment results.

Lemmatization

Also mentioned in Chapters 2 and 3, is the reported hypothesis that pre-trained contextualized language models have performed worse on Lexical Semantic Change Detection tasks than e.g. Word2Vec models because the test data was lemmatized while pre-trained models are 'used to' raw data (Schlechtweg et al., 2020). It is unrealistic given the results on the non-lemmatized data to verify whether pre-trained language models suffer from lemmatization of data, since no semantic variation has been detected in the first place. Still, it can be informative to test whether lemmatization has *any* effect on the measured semantic similarity between the target words. Therefore, in addition to performing the experiments on raw data, I also experimented with the lemmatized versions of the datasets. Before BERTje was fine-tuned and the representations were extracted, all comments in the datasets were lemmatized with the Dutch NLP pipeline from the open source library Spacy. Table 4.6 summarizes the results for the within-community and between-community comparison of representations of target words in lemmatized contexts, extracted from BERTje fine-tuned on the union of all lemmatized datasets. More detailed results are included in Appendix B.

Test data	Mean APS	
	Between	Within
Raw	0,78	0,78
Lemmatized	0,76	0,76

(*the Standard Deviation for each reported mean is 0,01*)

Table 4.6: Mean APS for representations extracted from raw and lemmatized test data

Lemmatizing the test data seems to have a small effect: the measured semantic similarity between representations of target words in lemmatized context is on average smaller than in raw contexts. A possible explanation for this result is similar to the one given for the observed difference between pre-trained and fine-tuned representations, which concerns the 'breadth' of the represented semantic information. With lemmatization, different word forms, each probably with slightly different meaning aspects, are counted as one word. Presumably, lemma representations relative to token representations therefore cover a wider semantic spectrum, resulting in greater semantic variation (and thus less similarity).

Hidden Layer Selection

In previous work on Lexical Semantic Change Detection methods using BERT to create word representations, different choices were made in which hidden layer(s) are extracted and possibly summarized into a word embedding. The different options did not seem to make a substantial difference in task performance. I aimed to verify this for the social semantic variation task. Instead of the concatenation of all hidden layers, only the top layer of BERTje was extracted as target word representation and the semantic variation experiments were also performed on these representations. The same pattern, as with lemmatizing the data or not fine-tuning the model, also seems to apply to a different hidden layer selection, visible in Table 4.7. The measured similarity between the representations formed by only the top hidden layer is lower than for the concatenation of all hidden layers, but again there is no difference between representations from the same and different social communities. The detailed results of these experiments, showing the APS per target word, have been added to Appendix B. The variation in APS between the individual words is slightly higher for the top layer representations, which is also reflected in the reported Standard Deviation. Still, no substantial community effect appears for any of the individual words.

Layer selection	Mean APS	
	Between	Within
All hidden layers	0,78 (0,01)	0,78 (0,01)
Top layer	0,73 (0,03)	0,73 (0,03)

(*Standard Deviation reported between brackets*)

Table 4.7: Mean APS for representations formed by all BERTje's hidden layers and only the top layer

4.3 Measuring Similarity

A sanity check to establish the validity of the measured similarity quantifications concerns the effectiveness of the metrics used. A failing metric could explain the fact that the method based on contextualized embeddings did not detect any community-dependent semantic variation. To verify whether the chosen metric is indeed functional, the measurements are not only performed between representations of the same word in different datasets, but also between representations of different words in different datasets. The results in Table 4.8 show the cosine similarity measurements between PPMI representations of different target words in different datasets. The values on the diagonal of the table (from top left to bottom right) show the results for the representations of the same target word in different communities. The latter measurements thus concern comparisons of representations that are semantically much more similar than the off-diagonal measurements. This is also confirmed by the metric, as the values on the diagonal are substantially higher. The cosine similarity metric therefore seems to be valuable to measure semantic similarity between PPMI word vectors.

CosSim		PS				
		klimaat	*media*	*belasting*	*overheid*	*liberaal*
FD1	*klimaat*	0,14	0,02	0,03	0,03	0,02
	media	0,01	0,11	0,02	0,03	0,03
	belasting	0,03	0,03	0,24	0,04	0,02
	overheid	0,03	0,03	0,05	0,17	0,03
	liberaal	0,01	0,03	0,02	0,02	0,06

Table 4.8: Cosine similarity between PPMI word vectors of different target words in different communities

The same testing procedure was applied to the Average Pairwise Similarity metric, which is used to measure similarity between contextualized word representations. The results in Table 4.9 also seem to confirm that the Average Pairwise Similarity metric is valuable to measure semantic similarity between contextualized word embeddings. A similar pattern to the cosine similarity check for the diagonal and off-diagonal values is visible, and even more apparent. The functioning of the APS metric therefore does not seem to explain the lack of detection of social semantic variation between the representations extracted by BERTje.

APS		PS				
		klimaat	*media*	*belasting*	*overheid*	*liberaal*
FD1	*klimaat*	0,78	0,35	0,39	0,36	0,34
	media	0,35	0,77	0,33	0,46	0,34
	belasting	0,40	0,33	0,76	0,42	0,35
	overheid	0,36	0,45	0,42	0,78	0,38
	liberaal	0,34	0,34	0,35	0,39	0,76

Table 4.9: Average Pairwise Similarity between contextualized word embeddings of different target words in different communities

4.4 Control Experiments

Comparing the similarity between the extracted word representations of the target words in datasets from different communities with those from the same community formed an important control setup to verify the community dependence of the measured semantic variation. This community dependency seems absent for the contextualized word representations, but not for the PPMI vectors. For the latter, an additional control experiment was performed to test the target word specificity of the detected social semnatic variation. The third and final control experiment addresses a possible explanation for the apparent absent social semantic variation for the contextualized representations.

4.4.1 Control Words

The PPMI-based method detected semantic variation which appears to be an effect of community difference. However, the question remains whether the observed difference in semantic similarity is specific for words that are conceptualized differently by users because of opposing political beliefs or whether this is, for example, a discourse-wide effect. To test this issue, the experimental set-up for the PPMI-based method has also been applied to a small set of control words that do not necessarily have a specific political affiliation. Table 4.10 shows for each of the control words the results of the cosine similarity between the PPMI vectors of the same word in different communities.

Control words	CosSim	
	Between	Within
boek	0,09	0,15
stad	0,09	0,16
stemmen	0,06	0,13
spreken	0,06	0,16
snel	0,04	0,09
nieuw	0,08	0,16
lopen	0,06	0,18
tafel	0,02	0,08
blauw	0,13	0,26
slapen	0,09	0,11
muur	0,12	0,19
warm	0,04	0,18
Average	0,07	0,15

Table 4.10: Cosine similarity between PPMI vectors of control words in different communities

The semantic similarity between the representations of the control words is overall lower than with the target words, which could be explained by the fact that the chosen control words may have broader meanings or be used in greater variety of contexts than the target words. For example, the control word *tafel* can be used in various figurative as well as literal contexts, as the examples below from the test data show:

> "*Laat ze dan de kaarten maar volledig op tafel leggen*"
> "*Als de discussie gaat leven onder de bevolking, krijg je gesprekken aan de keukentafel*"
> "*die problemen moeten vooral nu absoluut niet onder de tafel geschoven worden*"
> "*Tegen die tijd zijn ze blij als er nog eten op tafel komt.*"

More importantly, it turns out that also in this experiment there is a consistent difference between the within-community and between-communities comparisons. The cosine similarity between PPMI vectors in datasets from the same community is higher for all control words than between the vectors in datasets from different communities. This thus seems to indicate that the observed semantic variation does not exist specifically for the target words. It should be noted here that the difference in semantic similarity between the representations of different communities and the same community for some words, such as *blauw* and *warm*, is greater than for e.g. *slapen* and *tafel*. This implies that some words, such as the latter, can serve as stable control words. At the same time, it appears that selecting stable control words is hard as the degree of social semantic variation in the discourse of the two political communities also differs for words which in themselves do not directly indicate political relevance. The extent to which communities overlap and differ with regard to the conceptualization of language therefore still seems unclear.

4.4.2 Data Manipulation

Contextualized word representations do not show less semantic similarity for the same target word in different communities, while this is the case when using PPMI vectors as word meaning representations. As suggested by Schlechtweg et al. (2020), the lower performance of contextualized language models might be attributed to the influence of pre-training. More specifically, the possible explanation reads that the extracted word representations contain additional and effect-masking information due to pre-training. To test whether the context of a test instance has indeed (too) little impact on BERTje's word representations to detect semantic variation in the test data due to pre-training dominance, a manipulation experiment is performed. Following the strategy as described in Section 3.3.4, I have generated two extra datasets. Both are duplicates of the Forum_Democratie 1 dataset but a different manipulation was performed in each. In both duplicates, the context of a target word is manipulated by placing a target word, the donor, in the context of another target word, the recipient. If the contextualized representations of the donor word are indeed sensitive to context information from the test data, this should be apparent from changing token representations in the manipulated setting. Consequently, there should be a difference in semantic similarity measured for the donor word representations in the original dataset and the manipulated dataset.

In the first duplicate, the target word *klimaat* served as the recipient, and all it's mentions in the dataset were replaced by the donor target word *liberaal*. The original mentions of *liberaal* were replaced by the semantically 'empty' placeholder *foo*. This resulted in utterances where *liberaal* has been placed in unnatural contexts such as:

> "*Het standpunt omtrent liberaalverandering is beschamend.*"
> "*Wikipedia wordt geschreven door liberaalgekkies*"

Besides measuring the effect of this manipulation for the similarity between the representations of *liberaal* extracted from fine-tuned BERTje, the experiment was also performed with the pre-trained variant. This could provide insight into the extent to which the process of fine-tuning interacts with the context-specificity effect.

Table 4.11 shows the results for both of the experiments. The APS was measured between the token representations of *liberaal* in the Poldersocialisme dataset and the original and manipulated version of the Forum_Democratie 1 dataset. In the latter, the *liberaal* mentions are in the original contexts of the target word *klimaat*. If the context of a test instance has a major influence on the representation, it would be expected that the measured similarity with *liberaal* in the manipulated setting would correspond more towards the similarity with *klimaat* in the original setting, and thus would be much lower. This last measurement has therefore also been calculated. However, it can

Fine-tuned

APS		FD1		FD1_mnp
		liberaal	*klimaat*	*liberaal*
PS	*liberaal*	0,76	0,34	0,71

Pre-trained

APS		FD1		FD1_mnp
		liberaal	*klimaat*	*liberaal*
PS	*liberaal*	0,70	0,37	0,69

Table 4.11: APS between *liberaal* representations in PS and FD1_mnp (in which *klimaat* → *liberaal*), compared with *liberaal* and *klimaat* representations in the original FD1, extracted from fine-tuned (left) and pre-trained (right) BERTje.

be seen that the measured similarity with the representations of *liberaal* in the manipulated setting (0,71) is slightly lower than in the original setting (0,76), but not nearly as low as the similarity with *klimaat* in the original setting (0,34). This is even more extreme for the representations extracted from the pre-trained model, where the measured similarity with the representations of *liberaal* in the manipulated setting (0,69) is almost the same as in the original setting (0,70). So, the context of a test instance seems to have a little but almost negligible effect on the token representation of the target word. This supports the hypothesis that pre-trained information predominates in the final word representation of a contextualized language model. The difference between the results for fine-tuned and pre-trained BERTje seems to imply that fine-tuning on the test data mitigates the pre-training dominance effect somewhat.

Impact of Word Frequency in Fine-tuning

Given that the pre-training of BERTje has been done on billions of tokens and fine-tuning on millions of tokens seems to have a very small effect, an interesting question is: how much data does BERTje need to create contextualized representations which allow to detect any semantic variation present in that data? To take a first step towards answering this question, in the second duplicate of the Forum_Democratie 1 dataset the recipient word *links* is chosen, that occurs almost twice as often in the dataset (5025 times) than the recipient word *klimaat* in the first manipulation experiments (2839 times). All mentions of *links* in the second manipulated dataset were replaced by the donor target word *vaccinatie*. The original mentions of *links* were replaced by the semantically 'empty' placeholder *foo*. This resulted in utterances where *vaccinatie* has been placed in unnatural contexts such as:

> "*Het probleem met het neoliberale en vaccinatie gedachtegoed vind ik juist dat ze doorgeslagen zijn met sociaal individualisme*"
> "*Het is immers overduidelijk dat vaccinatie autoritaire, antidemocratische, pro-marxistische/communistisch trekjes hebben.*"

Table 4.12 shows the results of the experiments with the second manipulated dataset, which show almost exactly the same effects as the first. This implicates that doubling the number of usages to about 5000 does not seem to affect the apparent pre-training dominance.

The Most Contextual Layer

The data manipulation experiment is also performed with extracting only the top layer as word embedding, instead of the concatenation of all hidden layers. The results in Table 4.13 show for fine-tuned BERTje that the token representations of *vaccinatie*

Fine-tuned

	APS	FD1		FD1_mnp
		vaccinatie	*links*	*vaccinatie*
PS	*vaccinatie*	0,78	0,26	0,72

Pre-trained

	APS	FD1		FD1_mnp
		vaccinatie	*links*	*vaccinatie*
PS	*vaccinatie*	0,71	0,31	0,70

Table 4.12: APS between *vaccinatie* representations in PS and FD1_mnp (in which *links* → *vaccinatie*), compared with *vaccinatie* and *links* representations in the original FD1, extracted from fine-tuned (left) and pre-trained (right) BERTje.

in the Poldersocialisme dataset are more similar to the representations of *links* in the unmanipulated version, than to the representations of *vaccinatie* in the unmanipulated version. In manipulated dataset, the *vaccinatie* mentions are in the original contexts of the target word *links*, so this result was expected if the context of a test instance has a major influence on the representation. This was not the case in the previous experiment, with the representations formed by the concatenation of all hidden layers, thus confirming the outcome of the work of Ethayarajh (2019), which showed that the upper layers of contextualized language models are more context-specific than the lower layers. Still, previous experiments showed that no social semantic variation was detected even with the top layer representations. This seems to reject context insensitivity as an explanation for the non-detection, which I will discuss in more detail in subsequent chapters. Furthermore, pre-trained BERTje shows much less context sensitivity. This seems to be in line with the previous experiments implicating that fine-tuning on the test data mitigates the pre-training dominance effect.

Fine-tuned: top layer only

	APS	FD1		FD1_mnp
		vaccinatie	*links*	*vaccinatie*
PS	*vaccinatie*	0,71	0,43	0,50

Pre-trained: top layer only

	APS	FD1		FD1_mnp
		vaccinatie	*links*	*vaccinatie*
PS	*vaccinatie*	0,57	0,38	0,52

Table 4.13: APS between *vaccinatie* representations, formed by the top hidden layer only, in PS and FD1_mnp (in which *links* → *vaccinatie*), compared with *vaccinatie* and *links* representations in the original FD1, extracted from fine-tuned (left) and pre-trained (right) BERTje.

4.5 Leveraging Transparent PPMI-vectors for Fine-grained Insights

So far, the method based on PPMI representations has been shown to detect community-specific variation consistent with expectations. However, it has not yet been tested whether the way in which the representations vary also corresponds to the hypotheses formulated. Is the picked-up variation in, for example, the use of *klimaat* between the two communities indeed a matter of difference in conceptualization of *klimaat* as a problem versus hysteria?

Nearest Neighbor Analysis

For the purpose of analyzing the nature of semantic variation, the 10 nearest neighbors were computed for each target word based on cosine similarity between the PPMI vectors. In addition to the result *that* there is semantic variation for the target words between the different communities, this can provide more insight into *how* the communities differ semantically. The nearest neighbors of a target word give an indication of it's semantic interpretation within a community. However, they provide no information as to whether this interpretation really differs substantially from the other community. For example, if a word is the third nearest neighbor of a target word in one community and turns out to be the fourteenth nearest neighbor in the other community, this word is not really distinctive. To get more information about the degree of distinctiveness of a target word's nearest neighbor to a community, the position of this nearest neighbor in the ranking of the target word's neighbors in the other community has been calculated. I refer to this position as 'Rank'.

	Nearest neighbors of *klimaat* in			
Rank	Poldersocialisme		Forum_Democratie1	
	Lemma	Rank in FD1	Lemma	Rank in PS
1	verandering	1	verandering	1
2	**crisi**	162	akkoord	3
3	akkoord	2	wetenschapper	20
4	ecologisch	1215	**hysterie**	0
5	vestiging	0	ontkenner	8
6	**crisis**	1938	sverandering	0
7	**noodtoestand**	2535	plann	0
8	ontkenner	5	opwarming	13
9	kernenergie	21	milieu	10
10	milieu	9	verrandering	0

Table 4.14: 10 nearest neighbors of *klimaat*, based on cosine similarity between the PPMI-vectors, in different communities

To illustrate the outcomes of the nearest neighbor analyses, Tables 4.14 and 4.15 present the results for two of the target words that in the previous results (Section 4.1) showed a relatively larger difference between the within-community and between-community comparisons, i.e. *klimaat* and *kapitalisme*, respectively. In addition, Table 4.16 presents the nearest neighbors of *rechts*, a target word that showed a relatively smaller contrast. For all nearest neighbors analyses, the minimum occurrence frequency in the dataset for a word to be included was set to 10. A 'Rank' value of 0 indicates that the neighbor appears less than 10 times (or none at all) in the other dataset.

The nearest neighbors of the PPMI representations of the target word *klimaat* confirm the hypothesis that the members of `Poldersocialisme` interpret the concept as a problem, as evidenced by the nearest neighbors 'crisis' and 'noodtoestand', and that the members of `Forum_Democratie` interpret this concept as a hysteria, with literally 'hysterie' as one of the nearest neighbors. The high rankings or even absence of these neighbors (in bold in Table 4.14) in the other community's dataset also supports the fact that the interpretations really differ substantially. Nearest neighbors that appear in both communities' top ranks, such as 'verandering', 'akkoord' and 'milieu', could be interpreted as overlapping meaning components of the target word. These are explicable results as these words describe the sub-topics relevant to *klimaat* in the political domain. Based on the hypotheses, the nearest neighbor 'ontkenner' in both datasets is not an explicable result at first, because this concept does not seem to correspond with the interpretation of *klimaat* by `Poldersocialisme` members. However, a closer analysis of a data sample shows that in `Poldersocialisme` comments the 'ontkenner'-aspect usually has a negative relationship to *klimaat*, while this does not apply to `Forum_Demoratie` comments. The examples below illustrate this:

> "*Sommige ontkenners denken nog steeds dat ze klimaatverandering kunnen wegstemmen in het stemhokje.*" (Poldersocialisme)
> "*Misschien toch beter luisteren naar de klimaatontkenner.*" (Forum_Democratie)

The hypothesis regarding the target word *kapitalisme* mainly concerned a difference in connotation. `Forum_Democratie` members are generally positive about *kapitalisme*. The most distinctive nearest neighbors of the PPMI representation in the `Forum_Democratie` dataset appear, according to Table 4.15, to be terms for forms of *kapitalisme*, i.e. 'anarcho' and 'hyper', or inherent aspects of *kapitalisme*, i.e. 'hierarchie' and 'productiemiddel'. These neighbors do not necessarily carry a positive connotation, but also not a negative one.

	Nearest neighbors of *kapitalisme* in			
Rank	Poldersocialisme		Forum_Democratie1	
	Lemma	Rank in FD1	Lemma	Rank in PS
1	staats	90	socialisme	5
2	systeem	52	communisme	7
3	**imperialisme**	391	marx	194
4	**uitbuiting**	95	kapitalist	32
5	socialisme	1	anarcho	221
6	fundamenteel	1667	hiërarchie	1095
7	communisme	2	liberalisme	27
8	marxisme	25	productiemiddel	90
9	innovatie	1153	hyper	1813
10	**kut**	4114	isch	343

Table 4.15: 10 nearest neighbors of *kapitalisme*, based on cosine similarity between the PPMI-vectors, in different communities

The nearest neighbors of the PPMI representation in the `Poldersocialisme` dataset, however, do. This community was expected to have a negative association with *kapitalisme* and specifically view this economic system as exploitation. The nearest neigbors 'kut', 'imperialisme' and 'uitbuiting' and their ranking in the other community's dataset confirm these statements. The neighbor 'innovatie', which generally carries a positive connotation, seems at first to contradict the hypothesis, but closer analysis of a data sample reveals that this word often has a negative relation to *kapitalisme*, for example in:

> "*Innovatie komt niet door kapitalisme maar door investeringen*" (Poldersocialisme)

The other nearest neighbors, which apply to both datasets, are terms related to other political movements such as *socialisme* and *communisme*. The data of both communities also shows that economic or political systems are often contrasted in the discourse, thus appearing in similar contexts. An illustration of this is the following:

> "*Socialisme is echter het systeem dat als alternatief wordt geboden. Dus wat moet je doen? Een alternatief bieden. Lering trekken uit kapitalisme*" (Forum_Democratie)

The last nearest neighbor analysis I want to cover is that of the target word *rechts*. The cosine similarity results for the within-community and between-community comparisons showed that the difference between them was relatively smaller compared to the other target words. The nearest neighbors of the PPMI representation of *rechts* in Table 4.16 also confirm this observation, since almost all of them are not very distinctive from the other community.

	Nearest neighbors of *rechts* in			
Rank	Poldersocialisme		Forum_Democratie1	
	Lemma	rank in FD1	Lemma	rank in PS
1	extreem	1	extreem	1
2	treeks	5	taat	4
3	ut	8	praak	0
4	taat	2	conservatief	7
5	populistisch	51	treeks	2
6	radicaal	11	links	16
7	conservatief	4	ave	43
8	ing	8461	ut	3
9	extremistisch	13	alt	121
10	centrist	1734	zaak	54

Table 4.16: 10 nearest neighbors of *rechts*, based on cosine similarity between the PPMI-vectors, in different communities

Above all, it appears from the nearest neighbors that the less detected semantic variation could be explained by the fact that many of the neighbors are words that have nothing to do with *rechts* as a concept within the political domain, and thus the concept for which the hypotheses are formulated in terms of positive and negative connotations. These neighbors are parts of words that comprise the word form *rechts* and which have been incorrectly broken down through rule-based compound splitting, such as the words 'rechtstreeks' (directly), 'utrechts', 'rechtstaat' (rule of law), 'rechtspraak' (judciary) en 'averechts' (reverse). I will discuss the consequences of compound splitting further in Chapter 5.

PPMI-value Analysis

Since PPMI vectors are derivatives of co-occurence counts based on the word form of each lemma, with the application to semantic variation detection it is necessary to check whether the detected difference is really semantic in nature. Two synonymous context words differ from each other in word forms, but not in semantics. However, a count-based model makes no distinction and would thus detect two synonymous contexts as different. To determine whether the detected variation between the PPMI representations of the target words has been detected as variation for the right reasons, it is analyzed which values in the PPMI vectors, and thus which context words, differ most from each other for the same target word. Table 4.17 shows for each target word the top 10 context words corresponding to the PPMI values for which the difference between the two communities was greatest. There are two kinds of difference here: words that had a much greater value in the Poldersocialisme PPMI representation compared to the Forum_Democratie PPMI representation, and vice versa.

Target word	10 most distinguishing words (based on PPMI vector values) of	
	PS w.r.t FD1	FD1 w.r.t PS
klimaat	**noodtoestand**, onwaarheid, huizencrisi, ecologisch, pijpleiding, broeikasgas, parij, uitvoerbaar, tegenslag, **urgentie**	schommelen, hoofdoorzaak, **skeptisch**, toedoen, apocalyps, **gekken**, alarmistisch, hot, catastrofaal, **achterlijkheid**
vaccinatie	prober, zorgpersoneel, traag, sen, aller, achterhouden, gijzelen, verschijnsel, chip, wenden	pok, statussen, farmaceutisch, toedienen, ongevacineerd koorts, vind, organisator, incoming, uitrollen
immigratie	herkomst, creperen, homorecht, hekelen, watch, meerendeel, krimp, ontwikkelingshulp, hoefijzer, kampioen	innoveren, snood, sfinanci, medeverantwoordelijk, verregaand, gewog, sociaaldemocraat, nler, ongrondwettelijk, **funest**
vluchteling	**humaan**, oeigoers, maastricht, **begeleiden**, minsk, nls, deporteren, buurland, profiteur, leeuward	editie, goedzo, **kansarm**, maag, hekken, meevall, mileu, wonden, bedoeel, tiran
media	pipeline, uitgever, **reflectie**, storm, commerci, purpose, inlichtingendienst, netflix, journaal, misselijkmakend	trial, **fakenew**, oeps, consumeren, gepushed, informer, bevredigen, zondebok, verlengstuk, uitvergroten
belasting	procentueel, bbw, huiseigenaar, box, misdadig, begrotingstekort, periodiek, uber, terugwerken, fyi	verrekenen, belegger, **diefstal**, bregman, dwz, voila, katten, virtue, lonen, nalaten
overheid	pond, tienduizen, australisch, hernieuwbaar, actor, dictatorschap, mensbeeld, uitroei, volksgezondheid, voorwerp	gedragsverandering, libertarisme, corporatie, wob, bezuinig, ontelbaar, verlengstuk, postcodeloterij, regeldruk, bekostigen
links	seem, hobbie, terughoudend, versplintering, bakfiets, poepen, hoefijzer, memerij, authentiek, valkuil	kliek, **grappenmaker**, regressief, vooringenomenheid, profiteur, facistisch, idealistisch, zihni, nteerd, opschuiven
rechts	cancel, infiltrant, ing, **kutpartij**, rakker, **psychopaat**, honk, middenpartij, libertari, pijl	statelijk, tweestrijd, geori, nteerd, profiteur, schuwen, els, afbreuk, judas, nteeren
socialist	zadel, solution, scandinavi, **utopisch**, dsa, boekhandel, progressive, rotterdammer, bart, bolsjewistisch	hoezee, pur, entartet, singrijpen, omarm, orwell, manifesto, seat, herverdeling, **ugh**
liberaal	cuck, neutraliteit, stokpaard, arj, **bek**, progressivisme, neoliberal, wereldoren, champagne, potato	herverdeling, volkspartij, individualiteit, va, alom, mlk, centrist, thatcher, kaliber, liberal
kapitalisme	individualisme, afstappen, **boosdoener**, dictatorschap, nationaliser, entiteit, patriarchaat, bevel, managen, **bestelen**	hyper, **hoezee**, singrijpen, anarchistisch, resources, onstaan, aandeelhouder, dynamiek, sektarisch, krijg

Table 4.17: Words corresponding to 10 most distinguishing values in the PPMI vector of each target word in one community with respect to the PPMI vector in the other community

Words in bold seem to be in line with the hypotheses. What becomes clear from the words not in bold, is that the hypotheses of some target words are not really confirmed on the basis of the distinguishing words between the communities, e.g. for the target word *vaccinatie* and *overheid* in case of both communities, and for *immigratie* and *belasting* in Poldersocialisme. Furthermore, there appear to be distinguishing words that seem to contradict the hypotheses, as they are actually in line with the conceptualizations of the other community. However, closer analysis shows the hypotheses are actually not necessarily contradicted because these words appear to be used in combination with negation, sarcasm or a description of the opponent. For example, in the examples below from the Poldersocialisme subreddit, the underlined words, 'onwaarheden' and 'neutraliteit' by themselves seem to contradict the expected interpretation of the target words *klimaat* and *liberaal* respectively. However, it can be seen that 'onwaarheden' appear in an utterance describing an action by the political opponent Thierry Baudet. And about 'neutraliteit', its irony towards *liberaal* is literally made explicit.

> "*Thierry Baudet verkondigde in 2 minuten 18 onwaarheden over het klimaat.*"
> "*Dat is de ironie van 'liberale neutraliteit'*"

Examples that similarly invalidate the apparently contradictory distinguishing words can also be seen in the Forum_Democratie subreddit. In the examples below, the distinguishing word ‘catastrofaal’ in the context of *klimaat* is negated and ‘alarmistische’ is used in a description of political opponents. The latter also happens for ‘medeverantwoordelijk’ in the context of *immigratie*. In the last example, the cry of joy ‘hoezee’ is given in response to a *socialistische* state of affairs that is expressed with sarcasm.

> “*Het probleem voor de klimaathysterie ligt overigens niet bij de klimaatwetenschap. Het probleem ligt bij politici die zeer alarmistische uitspraken doen*”
> “*of Co2 uitstoot ons zo ver gaat brengen dat het ons klimaat catastrofaal veranderd, denk het niet*”
> “*Het probleem voor voornoemde partijen is dat (...) zij in feite medeverantwoordelijk zijn voor de uit de hand gelopen immigratie en integratie in ons land, met alle gevolgen van dien.*”
> “*De verdeling van welvaart is in ons oh zo socialistische Nederland nog nooit zo ongelijk geweest. Hoezee!*”

The cosine similarity measures between PPMI word representations in different datasets also revealed community-dependent semantic variation for control words. A refined analysis of the values in the PPMI vectors of the control words was therefore also performed. As for the target words, the top 10 most distinguishing words have been derived for each of the communities. These results have been added to Appendix B. Some example contexts related to these results are given below to illustrate that the distinguishing contexts for the control words provide explanations for the unexpected detected semantic variation. The control words, unlike target words like *overheid* and *immigratie*, were not expected to play a prominent role in the manifestation of different political beliefs. However, it appears that variation in the use of the control words, directly or less directly, can also be linked to different political views. The underlined word in each example is one of the ten most distinguishing context words of the target word present in the example for a community. In the `Poldersocialisme` subreddit, these context words, from the control words *tafel* and *blauw* respectively, contribute somehow to a description of typical leftist stances i.e. opposing the rich and standing up for the weak in society.

> “*Er word gestreden om broodkruimels van de tafel van de multinationals en miljardairs. Zet de kijkers op échte problemen.*”
> “*Voor advocaten is tegen de Belastingdienst procederen bijna liefdewerk oud papier, weet advocaat Khadija Bozia, die zeven cliënten met ‘blauwe problemen’ bijstaat.*”

In examples from the `Forum_Democratie` subreddit, including the same control words *tafel* and *blauw*, the distinguishing words also contribute to statements in which the views of the Forum voor Democratie party clearly emerge: fighting a mainly corrupt political system in which appearances are deceptive.

> “*Het rookgordijn is weg vriend, de waarheid ligt op tafel.*”
> “*Twitter in amerika word gedomineerd door extreem links. Hier in nederland lijkt het inmiddels ook al heel sterk links te leunen. Zeker alle blauwe vinkjes* omdat zij natuurlijk de beste deugmens willen zijn.*”
> *‘blauwe vinkjes’ refers to the blue verification badges that Twitter assigns to authentic famous persons such as politicians.

The cosine similarity results showed that the difference in semantic similarity between the representations of different communities and the same community for some words, such as *blauw*, is greater than for e.g. *tafel*. Still, from the examples of each of these words, manifestations of different political beliefs can be inferred. Thus, the extent to which communities differ in conceptualization of some control words appears to be greater than initially apparent from cosine similarity measurements between the word vectors. The broader scope of semantic variation over the discourse in this social setting is fully in line with the previously stated expectations of social scientist Mariken van der Velden, who predicted that the communities would talk differently about almost everything.

Chapter 5

Discussion

The experiments in this study mainly aimed to examine what information language models can provide about social semantic variation. Datasets from two different Dutch communities positioned at the extreme ends of the political spectrum reflected a social setting in which it has been theorized that opposing political views manifest themselves in different conceptualizations reflected by different usage. A list of target words for which this could apply with hypotheses about *how* the variation manifests, has been drawn up based on human expertise.

One of the key findings of the experiments is that word meaning representations created with a PPMI-based language model show hypothesized community-dependent semantic variation, while contextualized representations generated with the pre-trained transformer-based language model BERTje, either pre-trained or fine-tuned to the union of the test data, do not. Concerning the PPMI language model, it appears that the model detects not only variation between communities in representations of target words but also of control words. In addition, analyses of nearest neighbors and vector values of the PPMI representations can confirm more refined hypotheses about the nature of semantic variation, but also have shortcomings that warrant caution. BERTje, the contextualized language model, creates word embeddings that seem almost unaffected by the contextual information in the test data which could explain the poor performance. Fine-tuning mitigates the supposed pre-training dominance effect somewhat, but insignificantly.

In the following sections I will elaborate on the interpretations of the key findings and discuss the limitations of this research. I conclude this chapter with suggestions for future research arising from the discussed implications of this thesis.

5.1 Pre-training Dominance?

One of the most striking results of this study is the lack of detection of community-specific semantic variation by a state-of-the-art BERT-based language model where this is not the case for a traditional count-based model. This outcome goes against the current tendencies in the field of Natural Language Processing with pre-trained contextualized language models outperforming previous models on a wide variety of tasks. A possible cause for the discordant results, which also emerged in previous work on the detection of diachronic semantic variation (Schlechtweg et al., 2020), could be attributed to the pre-training information represented in these contextual embeddings. In contrast to count-based models that only incorporate information from the test data, the pre-trained transformer-based language models also contain a lot of pre-trained information. In fact, the amount of data fed to the model in the experiments of this thesis is only a fraction of the amount of data processed by the model during the pre-training phase. It is therefore speculated that the pre-training information represented in the word embeddings overshadows the test data information that would allow semantic variation to be reflected. With the simulation of semantic variation through data manipulation, an attempt has been made to investigate to what extent the context of the test instances influences the usage representations of the target words. To my knowledge, this approach for gaining more insight into contextualized representations has not been applied before.

In most of the tested setups, placing a target word in the contexts of a completely different word resulted in very small changes in the measured semantic variation between the communities for that target word. This seems to imply that the context in which the target word occurs has very little influence on the representation of the target word. The effect of the context manipulation turned out to be bigger for the fine-tuned model than for the pre-trained model, for which the effect was negligible. This difference would indicate that the process of fine-tuning results in representations that are slightly more sensitive to the contexts to which the model has been fine-tuned. This raises the question whether pre-trained models, after being exposed to sufficient data from the communities in question, would represent information in the word embeddings that allow to detect semantic variation between the communities. In other words, can sufficiently high-quality training data compensate for the apparent dominance of pre-training information? And subsequently, how much data would the model need for “strong enough signals”? A first experiment addressing these questions showed that doubling empirical data from about 2500 to about 5000 instances does not seem to affect the context-specificity. The question therefore remains whether this increase in quantity is trivial because of the much larger quantities that would be required or because an increased exposure to relevant empirical data does not have a linear effect at all on the context-specificity of the representations.

A factor that turned out to have a major influence on the context-specificity of the representations is the selection of the hidden layer(s). When only the top hidden layer in fine-tuned BERTje was used as word representation, placing a target word in unnatural contexts resulted in a substantial similarity drop. The top hidden layer thus seems to be a lot more context sensitive than the concatenation of all hidden layers, which confirms the findings of Ethayarajh (2019). However, this difference in hidden layer selection did not appear for pre-trained BERTje. Moreover, the difference in the hidden layer selection (also for the fine-tuned variant) seemed to have no effect on the main results of this study. When only the top layer, i.e. the relatively most context-specific layer, was taken as word representation, no community-specific semantic variation was detected either. This result seems to contradict the speculated correlation between pre-training dominance or context sensitivity and the detection of semantic variation, and thus the possible explanation for the lower performance of pre-trained contextualized language models on semantic variation tasks.

5.2 Extend of Social Semantic Variation

The PPMI model did detect the predicted social semantic variation, but also the unpredicted variation, i.e. variation for control words that are not necessarily fundamental political concepts. Although it was more convincing for some control words than for others, the similarity between representations of all control words was lower between datasets from different communities than from the same community. A failing method could have been an explanation. However, this result was roughly predicted by the political communication expert Mariken van der Velden. In addition, explanatory PPMI-value analysis together with examples from the datasets showed that the detection of semantic variation for the control words does not appear to be invalid. In addition to the target words, i.e. prominent political concepts, control words also appeared to play a role in expressing different political beliefs.

At least two implications relevant to different scientific disciplines can be drawn from what appears to be a wide extent of semantic variation in social settings. For the social sciences, and more specifically the novel field of Cognitive Sociolinguistics, these results contribute to new insights into recent social phenomena. Since Van der Velden was not yet aware of any existing literature that motivates her predictions about the discourse-wide variation between social communities, the results in this thesis seem to be the first evidence-based indications for this. More specifically, the results seem to indicate that different interpretations of word meaning by different political communities do not only apply to fundamentally political concepts, but manifests itself much more widely across the discourse. However, the indications are based on a limited selection of target and control words, so drawing firm conclusions therefore requires larger-scale analyses.

The second implication, which follows from the first, is more relevant to the field of computational linguistics and concerns the application of computational methods for the analysis of *temporal* semantic variation to *social* semantic variation. The validity of this principle, on which the studies of Del Tredici and Fernández (2017) and Lucy and Bamman (2021) also build, calls for a critical reconsideration as the two phenomena seem to have different characteristics. Diachronic semantic variation mainly manifests itself in a gradual shift of the meaning of a particular word. The process of the shift usually includes a period in which the original and the new word meanings coexist (Traugott, 2017). Synchronous semantic variation, such as between the social communities studied in this research, seems to extend more widely across the discourse, and does not just apply to a single word. In addition, the different meanings seem to be almost mutually exclusive. Since the different forms of semantic variation thus appear to be expressed differently, it is not necessarily justified to treat them equally in computational detection.

For example, the selection of control words for testing computational detection of *social* semantic variation seems more difficult than for *temporal* variation. The experiments in this thesis showed that words that seemed stable control words in themselves, showed community differences in closer analysis. A recommendation that follows for future research is that the selection of control words may require manual corpus analysis, as was done for the target words.

5.3 What do PPMI-vectors Tell?

In addition to presenting a method based on PPMI representations to detect *that* social semantic variation exists, this thesis also demonstrated how this type of word representations can be employed to analyze *how* word meanings vary. These strategies can provide more refined insights, while simultaneously testing whether the computational detection of social semantic variation is for the right reasons. A possible ‘wrong’ reason for detected variation to be tested could be a non-semantic nature of variation because the context words determining the difference in word representation are synonyms of each other. This unwanted corpus variation is a realistic scenario in count-based models as the vectors are based on word form only.

The applied strategies for obtaining more refined insights include the comparative analyses of 1) the nearest neighbors of the community-specific PPMI vectors and 2) the values in the PPMI vector (directly corresponding to context words). The results proved to confirm specific hypotheses about the conceptualization of the target words in the different communities and provided explanations for the detection of semantic variation in the control words. However, the fruitful possibilities of transparent PPMI representations also have limitations. PPMI representations of word meaning, like other representations resulting from distributional methods, are not sensitive to negation or sarcasm in the contexts of the words. As a result, words with vectors equal to the target word, i.e. the nearest neighbors, or the context words that are distinctive for a target word in a particular community, do not certainly contribute to a correct reflection of the community-specific interpretation of the target word. For example, when using the context word 'innovatie' only in negative relation to the concept of *kapitalisme*, the analysis of the PPMI representations indicates a strong association between the words 'innovatie' and *kapitalisme*. Drawing incorrect conclusions is therefore a risk for which manual exploration of the test data can offer a solution. The results in this thesis therefore show on the one hand how the analysis of PPMI representations can provide refined insights, but also show its dangers that should be dealt with caution.

5.4 Limitations of this Research

This study has some limitations that should be taken into account when interpreting the results. First, I will discuss the main caveat, which concerns the limitations in using language use as a tool to study social semantic phenomena. I then discuss the consequences of the methodical component of compound splitting. Finally, the setting of hyperparameters is criticized for the exhaustion of the tested capabilities of language models.

What does Language Use Reflect?

The Distributional Hypothesis provides a reasonable framework for analyzing word meaning. However, for the analysis of social semantic variation, two caveats can be made when using this framework regarding 1) the underlying factors for different language use and 2) the conclusions that can be drawn from different language use. As mentioned briefly before, the effect that different language use is caused by different discourse topics should be excluded in the analysis of semantic variation in the setting of this thesis. For the selection of target words, this alternative underlying factor for different language use between the two politically-oriented communities is largely excluded. That is, if the target words are used, it is in the context of similar political topics. This assumption generally held, but there were also exceptions. Closer analysis of the data showed, for example, that the target word *rechts* is also used for legal purposes. However, a more critical comment regarding the discourse-topic effect can be made in the experiments with control words. To test whether the detected semantic variation also applied to concepts other than fundamentally political concepts, control words were chosen that had minimal political significance, but were fairly frequent in discourses of both communities. This resulted in a selection of control words with more ambiguous meanings, thus increasing the possible variation in discourse topics in which these words are used. Although analysis showed that the contexts comprising the control words generally discussed topics related to politics, these topics differed several times in the two communities. For example, the control word *blauw* in the `Polder-socialisme` subreddit was used in the context of tax issues (where blue refers to the color of the envelopes from the Dutch tax authorities), while in the `Forum_Democratie` subreddit *blauw* was used in the context of the verification badges on Twitter, and thus media related. In this case, the difference in language use is not necessarily caused by different meanings of *blauw*, but rather because the word occurs in contexts related to different topics. Observations like these give reason to question the conclusions drawn from the results for the control words, since the discourse topic as an underlying factor for different language use does not necessarily appear to be excluded.

The second concern regarding the use of the Distributional Hypothesis for the analysis of semantic variation relates to the distinction between word usage and word interpretation. In the current field of computational analysis of semantic phenomena, a word vector is usually interpreted as the representation of word's meaning. Here, the tendency to interpret a word vector as the user's interpretation of the word is tempting, but not necessarily justified since the vector is only based on the use of the word and not on perception information. With regard to the research objectives, it must therefore be noted that the results of a computational method based on the Distributional Hypothesis can provide answers to a limited extent. Specifically for the outcomes in this thesis, the critical question can be asked whether talking about a political concept in a particular way is a reflection of a community member's sole interpretation of the concept or whether there is a broader interpretation that is not fully expressed? If the latter is the case, a difference in computational word representations would not provide isolated evidence for the manifestation of political polarization in the word meaning of the concepts tested. Computational analysis of semantic phenomena using the current framework can thus provide sound and valuable indications, but not hard evidence with regard to language conceptualization.

Consequences of Compound Splitting

One of the design choices in the used methods was the splitting of compounds. The Dutch language is characterized by a high degree of compound formation. If compounds consisting of one of the target words were left in original form, not only would this result in a substantially lower amount of empirical data for each target word, but above all a lot of valuable information would be ignored, which would affect the representativeness of the results. For example, the contrast between the compounds 'overheidssteun' and 'overheidsbemoeienis' is an explicit reflection of the different meaning interpretations of the target concept *overheid*. This would have been missed if only words separated from the target word were included as context.

However, the way the compounds are split in the experiments has some pitfalls. Using a simple rule-based strategy, spaces were inserted before and after each mention of a target word to separate the target word from the other compound parts. This also separated words that contain the word form of the target word but are not compounds at all. For example, the adjectival derivation of the Dutch city of Utrecht, i.e. 'utrechts', includes the word form of the target word *rechts* but is not a compound word. As a result some “false” mentions of a target word have been caught, bringing along contexts that detract from the representativeness of the resulting word vector. In addition, the pre-separation of compound words interferes with the subword tokenization of BERTje. BERTje distinguishes the first sub-word from the subsequent subwords by means of the ## symbol. For example, when the word ‘vluchtelingenstroom’ is split into subwords by BERTje’s tokenizer, this results in:

> ‘[CLS]’, ‘vluchtelingen’, ‘##stroom’, ‘[SEP]’

with the ## symbol indicating that ’stroom’ is a subsequent subword. Moreover, this result also shows that ’vluchtelingen’ is not split into two subwords, i.e. the singular form ’vluchteling’ and the plural morpheme ’en’. The target word would thus not be captured here with the default tokenization. Pre-splitting of the compounds comprising a target word does result in the target word being separated from the rest allowing it to be captured. In the case of ’vluchtelingenstroom’ the result after BERTje’s tokenization is then as follows:

> ‘[CLS]’, ‘vluchteling’, ‘en’, ‘##stroom’, ‘[SEP]’

However, the modification before BERTje’s tokenization does have the consequence that BERTje identifies not only ‘vluchteling’ but also ‘en’ as the first subword. So far it is unknown to me to what extent the identification of subsequent subwords as opposed to first subwords influences the predictions of the model. In order to strengthen the validity of the results in this study, follow-up research should rule out whether this difference in context representation has an effect on BERTje’s performance.

Hyperparameter Settings

Finally, the possibilities of the tested models could have been exploited even better if more experiments had been done with the values of the hyperparameters. A fixed window size of 10 tokens has been taken for the implementation of the PPMI model. No context length other than the maximum possible context length of the contextualized model, i.e. 512 tokens, has been tried for the configuration of BERTje. For fine-tuning, the default settings of the Hugging Face implementation are used for the values of training hyperparameters such as learning rate. Different setting of these values could lead to better or worse model performances that could influence the conclusions of the experiments.

Chapter 6

Conclusion

6.1 Summary

This thesis has contributed to the need for more insight into the growing political polarization manifesting in language. With a multi-disciplinary approach I explored the computational linguistic possibilities for the analysis of social semantic variation. I have combined studies into the use of language models for detecting shifts in word meaning, i.e. Lexical Semantic Change Detection (LSCD), with the new research field of Cognitive Sociolinguistics. Studying semantic variation from a socio-cognitive perspective is not only more relevant in an age of online communities, but also more accessible than ever with the enormous amounts of data on social media. Harnessing the data accessibility and addressing the need for greater insight, this research exploited the potential of using language models to analyze semantic variation in a social setting, with specific attention to the recent developments in contextualized language models. All in all, the two-fold research issue focused on 1) the information that language models are able to confirm about cognitive-linguistically informed hypotheses and 2) the difference in performance between a state-of-the-art transformer-based model and a more traditional count-based model. I addressed these issues by selecting and applying two LSCD methods to the analysis of semantic variation between two Dutch-speaking social communities on Reddit that differ fundamentally in political affiliation.

Strong emphasis of this research is on the evaluation of the computational methods. A combination of different evaluation procedures have been applied to enhance the validity and sophistication of the conclusions. The main evaluation concerned the testing of the methods on a list of predefined target words for which concrete hypotheses exist regarding conceptual variation between the communities. These hypotheses were based on both a manual corpus analysis and the expertise of a social scientist on political communication.

The methods tested included 1) the creation of meaning representations for each target word in both communities and 2) the measurement of the similarity between the community-specific representations of each target word. In the count-based method, PPMI representations were created and compared using the cosine similarity between them. The method based on a contextualized language model extracted the hidden layers of the Dutch pre-trained transformer-based language model BERTje and compared these contextualized representations between the different communities with the Average Pairwise Similarity measure. To validate whether the measured similarity between representations of the same target word in different communities reflected social semantic variation, different control setups were used.

The comparison of representations in different datasets of *the same* community with those of *different* communities showed that the PPMI model does reflect community-dependent semantic variation, while BERTje does not. This is a striking result given that pre-trained contextualized language models substantially outperform more traditional models on many NLP tasks. One of the possible reasons for this result is that context information represented in the word embeddings is overshadowed by pre-training information. From experiments with context manipulation set-ups, BERTje's top hidden layer, in contrast to the concatenation of all hidden layers, seems to be sensitive to the context of the test data if the model is fine-tuned on this data. However, the top layer representations also do not detect social semantic variation. Experiments testing the influence of fine-tuning and lemmatization showed no effect on the variation detection either. Although the different configurations and data manipulations in the experiments did not provide an explanation, the observation remains that pre-trained contextualized representations for the tested set-ups do not seem to represent the information reflecting social semantic variation. This could imply that a pre-trained language model is not always the most successful choice which would have consequences for e.g. social sciences leveraging these state-of-the-art models with the belief that BERT-like models are the most suitable due to the dominating high-performance tendencies.

The PPMI vectors, by contrast, not only reflected social semantic variation but also proved useful for more refined testing of the hypotheses because of their transparent nature. Analysis of the nearest neighbors and community distinguishing values of PPMI vectors confirmed several hypotheses about *how* the conceptualization of target words varies, but also showed that caution should be exercised in this interpretation due to the limitations of a bag-of-words approach. Furthermore, following the detection experiments with *control words*, the fine-grained analyses offered more well-founded indications that different interpretations of word meaning by different political communities do not only apply to fundamentally political concepts, but manifests itself much more widely across the discourse.

6.2 Future Work

The work in this thesis offers new insights, but the results and limitations of this research also raise the need for follow-up research. First of all, the last observation in the summary above, which is also discussed in more detail in Chapter 5, points to a distinction between social and temporal semantic variation. Synchronic semantic variation, as studied in this research, seems to extend more widely across the discourse and does not just apply to a single word. These findings are based on a fairly small-scale analysis, but its implications would be of great importance. Human assessments of word interpretations in different social contexts could provide more insight relevant not only to political and social sciences but also to the development of computational methods that may or may not be used. So, larger-scale research in which the properties of social semantic variation would be further described is needed. In addition, more expertise on human interpretation can also provide more information about what can be concluded on the basis of language use. Specifically, greater clarity on the extent to which the reflection of language use corresponds to perception could validate the reliance on the Distributional Hypotheses for conclusions regarding language manifested polarization.

Second, the results in this thesis call for more research into the use of pre-trained contextualized language models. The underperformance of this type of model on the social semantic variation detection task could possibly be explained by the fact that the context information from the test data is overshadowed by pre-training information in the hidden layer representations. This statement seemed to be confirmed in many but not all configurations. Representations consisting of the top hidden layer of the model fine-tuned on the test data were clearly sensitive to manipulations in the test data. A question that has not yet been fully answered in this thesis is whether the amount of data on which BERTje is fine-tuned has an effect on the degree of context specificity of the representations. A first contribution to this question was made by testing a double amount of mentions of a target word (namely 5000), which had no effect. Follow-up research should rule out whether there is indeed no relationship between the size of the fine-tuning data and the context-specificity of the representation, or whether that relationship does exist but only becomes visible with much larger quantities.

Still, no community-specific semantic variation was detected even in the configuration with the most context-sensitive representations. This configuration included fine-tuning the model. A possible explanation for the lack of detection could be that the model has been fine-tuned to the *union* of test data from both communities. The fine-tuning information that would be represented in the word embeddings is therefore not community-specific. In follow-up research, a setup could be tested in which the contextualized language model is fine-tuned separately on the dataset of each community. Such a setup requires a post-hoc alignment technique to be applied, as in earlier studies of e.g. Hamilton et al. (2016b) who used a Word2Vec model. In addition, the influence of pre-training information could be completely excluded by taking a contextualized model that is not pre-trained on any data other than the test data, such as ELMo (Peters et al., 2018). The downside is that such a model needs much more data than in the current study to generate effective word representations (Sahlgren and Lenci, 2016). In previous studies on Lexical Semantic Change Detection, the predictive models that generate static embeddings, which also require a lot of data, turned out to provide the best results (Schlechtweg et al., 2020). It would therefore be interesting to see whether the contextualized variants, because of the richer representations they can create, could perform better analysis of semantic variation without the apparently disruptive pre-training or shared fine-tuning information.

Appendix A

Data examples per topic

Topic	Examples	
	Forum_Democratie	**Poldersocialisme**
Climate	*klimaatverandering is een hoax*	*in een tijd waarin we met zijn allen zouden moeten vechten om het klimaat te redden*
	blinde klimaatgekte	*als t klimaat een bank was dan was het al gered*
	het probleem is echter dat het overwegend niet de weldenkende nederlanders zijn die deze klimaathysterie voorstaan	*zolang kapitalisme buitensporige consumentencultuur blijft creëren dan zijn alle klimaatmaatregelen dweilen met de kraan open*
	uiteindelijk gaat het erom wat goed is voor nederland, mensen van het klimaatgeloof af krijgen dus	*het kapitalisme is gebouwd op eeuwige groei die geforceerde groei heeft het klimaatprobleem met zich meegebracht*
	klimaat hysterie dat alleen maar onze agrarische sector kapotmaakt	*miljarden uitgeven aan het klimaat gaat zorgen dat mensen het echt beter krijgen*
Covid-19 vaccination	*omdat iedereen vrij zou moeten zijn of je vaccineert of niet*	*vaccines verkleinen de kans op verspreiding met 97 procent*
	net als dat het niet aan een overheid is om iemand te dwingen een vaccinatie te nemen	*vrijwel alle opgenomen mensen in de ic zijn ook niet gevaccineerd*
	de vaccinatiepaspoorten waar we mee te maken krijgen dat zijn echt wel vrijehidssbeperkende maatregelen waarvan het het gevaar bestaat dat ze blijvend zijn	*met vaccinaties kunnen we zoveel mogelijk voorkomen dat het virus naar gevoelige groepen verspreid*
	het begint met vaccinaties maar waar eindigt het als het verplicht word verlies je het recht op de keuze wat jij met jouw eigen lichaam wilt kunnen doen	*mensen die vaak een slecht immuunsysteem hebben zijn juist het grootste slachtoffer van de egoïsten die weigeren gevaccineerd te worden*
	ik vind dat de individuele vrijheden boven het collectief belang staat daarom geen enkele directe of indirecte vaccinatieplicht wat mij betreft	*niet vaccineren belemmert de vrijheid van het collectief*
Immigration, refugees	*het is de enige partij die realistische en goede plannen heeft de massaimmigratie te stoppen*	*de beste manier om de culturele diversiteit van vluchtelingen te behouden*
	echte problemen zoals illegale immigratie, toenemende criminaliteit en afnemende veiligheid	*omdat vluchtelingen niet voor niets vluchten en wij als rijk land zeker verantwoordelijkheid voor andere dragen*
	ik wil dat de immigratiekraan dichtgedraaid wordt	*migranten zijn vaak vluchtelingen, mensen die vreselijke omstandigheden ontvluchten*
	sinds een jaar of vijftig geleden is door corrupte politici een massale stroom immigranten binnengelaten waardoor de hele samenleving en straatbeeld ermee is verandert en dat heeft veel verdeling en een toename aan criminaliteit veroorzaakt	*mijn hart breekt gewoon als ik de vluchtelingen in onmenselijke sporthallen zie*
	met name de recente vluchtelingencrisis staat onze cultuur nog meer onder druk	*de arbeidsmigrant aanwijzen als schuldige is uiteraard verkeerd, de schuld ligt bij brusselse regelgeving*
Media	*niets van geloven, is allemaal frame van de media*	*het delen van informatie en media is juist goed voor de maatschappij*
	waarna de media alles opblaast	*mijn mening op basis van de berichten in de media was dat*
	politiek is een grote poppenkast met de media als oorlogswapen aan hun zijde	*ik zie regelmatig om mij heen en ook in de media dat veel mensen het niet zo breed hebben*
	ondanks alle misinformatie vd media de laatste jaren	*de media heeft er alles aan gedaan om dit te normaliseren*
	de media schreeuwen dus al een week moord en brand zonder fundering	*ik zou gewoon de mainstream media volgen die hebben namelijk objectieve informatie*
Taxes	*we worden al links en rechts bestolen door belasting*	*laten we inderdaad maar eens beginnen met flinke kapitaalbelastingen*
	zou de inkomensbelasting naar beneden kunnen zodat werken blijft lonen	*ga liever meer belasting heffen op milieuvervuiling of op idioot hoge vermogens*
	in ieder geval geen belastingverhoging of straffende kapitaalbelasting	*als we daarvoor belachelijk hoge belastingen moeten invoeren op hoge lonen vind ik dat helemaal prima*
	belasting is diefstal	*je kan beter de belastingen op multinationals omhoog gooien*
	sorry maar de staat steelt van ons, belasting heffen en dan lekker kansloze immigranten gaan huisvesten	*om te beginnen een progressieve belasting op vermogen*

Table A.1: Examples from the dataset illustrating the motivation for the hypotheses formulated about the conceptualization of the target words by the different communities (as presented in Chapter 3). Part 1 out of 2.

Topic	**Examples**	
	Forum_Democratie	**Poldersocialisme**
Government	*de overheid bestaat niet om mij te helpen het bestaat om mijn vrijheid te waarborgen*	*daarom is een sterke overheid in mijn ogen een eis om de burger te beschermen tegen de werkgever*
	ik geloof dat het de taak van de mens en niet de overheid is om uit zichzelf te geven om de naaste	*tijd voor de overheid om in te grijpen*
	uitgaven van onze overheid weer zelf gaan bepalen	*al lang blij dat de overheid wat doet aan de radicalen daar*
	de overheid moet zich zo min mogelijk bemoeien met de persoonlijke levenssfeer	*de overheid had veel meer de regie moeten houden voor voldoende woningen*
	het doel is om de belasting te verlagen en op die manier de grootte van de overheid ook te verlagen	*ben ik sterk van mening dat de overheid zich meer met de gevolgen van automatisering moet bemoeien*
Political movements, economic system	*krijg je opeens een random domme zure linkse opmerking naar je toe geslingerd*	*de feiten geven best duidelijk aan dat links op zo goed als alle fronten gelijk heeft*
	ook hier is het gevaar van links te merken	*revolutie naar links naar meer gelijkheid en dat willen we nog steeds*
	de basisprincipes van links slaan compleet nergens op	*een groot deel van de samenleving is niet links helaas*
	de media doet nu alsof rechts gelijk extreemrechts is, maar met rechts en nationalistisch zijn is niks mis	*een maatschappij die zich nog makkelijker dan nu zal laten bespelen door de populisten op rechts*
	dacht ik eindelijk een normaal rechts artikel te lezen	*de onmenselijke politiek van rechts*
	mensen die roepen dat rechts het probleem is	*die psychopaten van fiscaal conservatieve rechtse partijen*
	vergiffenis is een concept wat ze niet kennen onder de socialisten en de communisten	*het doel van de meeste moderne socialisten in de 21ste eeuw is de economie vermaatschappelijken*
	gadver socialisten	*dat is niet het electoraat dat je wilt als solidaire socialisten*
	socialisme is slecht	*dat we ons als socialistische beweging moeten verzetten tegen de echte oorsprong van consumptiecultuur*
	nederland wordt opgeslokt door een socialistisch instituut soevereigniteit wordt afgenomen	*we maken duidelijk dat we een partij zijn die volledig voor het socialisme strijden*
	de grondlegger van het fascisme was een keiharde socialist	*socialisme is een zo veel opzichten een juiste stap maar wat landen als nederland zo groot heeft gemaakt*
	overheidsbemoeienis via liberalisme stuk minder	*de rotzooi die veroorzaakt is door neoliberaal beleid*
	in de economie zorgt liberalisme voor ongelijkheid maar ook voor gigantische onevenaarbare welvaartsgroei de voordelen zijn groter dan de nadelen	*wat het betekent om in een kwetsbare groep te zitten en je leven op pauze te hebben dan zijn de scherpe randen van het neoliberalisme heel erg voelbaar*
	een ware liberaal zal en moet zijn vrijheden verdedigen	*als die overheid winsten privatiseert en schulden socialiseert is het neoliberalisme*
	ik ben een conservatieve liberaal minder belastingen en meer eigen verantwoordelijkheid	*objectief en empirisch bewezen is dat neoliberalisme aan aanslag is op duurzaamheid van onze samenleving*
	het mooie van het liberalisme is dat niemand de politieke wil van een ander opgedrukt krijgt	*liberalisme hangt met leugens aan elkaar*
	dankzij het kapitalisme heb jij zelf de mogelijkheid om het kapitaal te genereren om de middelen voor het vinden van kennis te bemachtigen	*partijen die het afschuwelijke kapitalisme in stand willen houden*
	kapitalisme is uiteindelijk vaak de oplossing	*kapitalisme synoniem voor bittere armoede uitbuiting en oorlog*
	met kapitalisme maar dan echte kapitalisme genereer je de meeste welvaart	*kapitalisme is niet compatibel met de menselijke aard*
	ik zou zeggen dat kapitalisme een geweldige verbetering vormde op de voorgaande en huidige alternatieve systemen	*een slaaf van de winsteisen van kapitalisten*
	vaak krijgt kapitalisme ook de schuld van overheidsingrijpen	*de kapitalisten zijn het probleem, niet de overheid*

Table A.2: Examples from the dataset illustrating the motivation for the hypotheses formulated about the conceptualization of the target words by the different communities (as presented in Chapter 3). Part 2 out of 2

Appendix B

Results extra experiments

Target words	APS between		
	FD1&PS	FD1&FD2	FD2&PS
klimaat	0,78	0,78	0,78
vaccinatie	0,78	0,79	0,78
immigratie	0,79	0,79	0,79
vluchteling	0,80	0,80	0,80
media	0,77	0,77	0,77
belasting	0,76	0,76	0,76
overheid	0,78	0,78	0,78
links	0,77	0,77	0,77
rechts	0,75	0,75	0,75
socialist	0,78	0,78	0,78
liberaal	0,76	0,76	0,76
kapitalisme	0,79	0,78	0,79
Average	0,78	0,78	0,78

Table B.1: Average Pairwise Similarity between token representations of target words in all different datasets

Target words **not** in	Fine-tuned		Pre-trained	
BERTje's vocabulary	Between	Within	Between	Within
vaccineren	0,73	0,73	0,63	0,63
gevaccineerd	0,72	0,73	0,58	0,60
immigrant	0,75	0,75	0,62	0,61
socialisme	0,75	0,75	0,61	0,60
liberalisme	0,72	0,72	0,60	0,58
kapitalist	0,74	0,74	0,59	0,58
Average	0,74	0,74	0,60	0,60

Table B.2: Average Pairwise Similarity between token representations of target words that did not pre-exist in BERTje's vocabulary

Target words	Lemmatized		Pre-trained		Top layer	
	Between	Within	Between	Within	Between	Within
klimaat	0,77	0,77	0,70	0,70	0,75	0,76
vaccinatie	0,77	0,78	0,71	0,71	0,71	0,73
immigratie	0,76	0,76	0,72	0,72	0,73	0,73
vluchteling	0,76	0,76	0,73	0,73	0,76	0,76
media	0,75	0,76	0,71	0,72	0,71	0,73
belasting	0,73	0,74	0,71	0,72	0,69	0,69
overheid	0,75	0,76	0,70	0,70	0,77	0,77
links	0,75	0,75	0,72	0,72	0,75	0,75
rechts	0,73	0,73	0,69	0,69	0,70	0,70
socialist	0,77	0,76	0,71	0,70	0,75	0,74
liberaal	0,76	0,76	0,70	0,70	0,68	0,70
kapitalisme	0,77	0,76	0,70	0,69	0,75	0,74
Average	0,76	0,76	0,71	0,71	0,73	0,73

Table B.3: Average Pairwise Similarity between contextualized representations with different settings (if not mentioned otherwise: fine-tuned BERTje, non-lemmatized data, all hidden layers extracted)

Control word	10 most distinguishing words (based on PPMI vector values) of	
	PS w.r.t FD1	FD1 w.r.t PS
boek	bolsjewistisch, boekhandel, understanding, farm, pannekoek, jori, damage, overzicht, scheidslijn, leesbaar	signeren, schwab, arnold, honken, uiteen, handmatig, voorwaardelijk, detailleren, omg, overtrekken
stad	verderop, hoogbouw, opmars, integreren, briljant, filiaal, versus, zwolle, pikachu, britten	jeruzalem, shithole, gemeenteraadslid, jag, rivier, gebruikersnaam, bombardement, lond, luchtkwaliteit, hinten
stemmen	bedenktijd, respectabel, democraten, kutt, christenunie, aziat, ondersteuningsverklaring, eurosceptisch, petra, wybr	remain, overweg, rep, kiesman, ehhh, jee, asielaanvraag, doorslag, zionistisch, wegtrekken
spreken	lam, zorgverlener, juf, betrokkenheid, watch, nterviewd, geliefd, meta, ba, griek	foutloos, knul, familielid, abn, accent, leeuward, tjah, hoeksteen, vrijuit, inburgeren
snel	telling, bloedgroep, gebasseerd, yike, spits, uitwijzen, vaart, fort, radioactief, monopolist	scheen, koolhydraat, eindig, tij, goedzo, smelt, detecteren, aggressief, voornaam, owens
nieuw	autos, pensioenstelsel, nav, royer, komaf, bespelen, stelsel, beat, uitkeringsgerechtigd, spaarpot	testament, opduiken, oof, toestel, teleurstelling, redacteur, zuil, postmodernisme, introduceren, fdf
lopen	eend, rotterdammer, kap, overwerken, vleesindustrie, tag, tegenaan, menigte, graaf, sketch	raaskallen, hardlopen, vertraging, redditors, gaap, ingang, hoesten, uitzonderen, levensgevaar, onveiligheid
tafel	teringzooi, nimmer, erkenbrand, taart, kruimel, boterham, gvd, grappen, achterblijven, eenmalig	vuist, facepalm, wob, huys, jovd, memo, biljet, gordijn, achja, kenner
blauw	kaas, cli, nten, stapel, bernard, universum, ogenschijnlijk, wuiven, bijstaan, puber	woonkamer, verven, femini, thumbnail, network, vink, zegmaar, leesbaar, smaken, parafrase
slapen	hotel, thee, voeg, eigelijk, wee, tandenborstel, dijsselbloem, parkeerplaats, opstarten, vernieling	spuitje, kussen, bedragen, dromen, tweedeling, stef, huisdier, gezelschap, geweten, vernieling
muur	linkiewinkie, verwoording, character, ongein, overwinnen, analogie, verven, twaalf, overstap, gelijkstemmen	ruyter, sikkel, vlieg, poort, tegenop, slacht, schilderij, mussert, patriottisch, automatiseren
warm	trui, maaltijd, everyone, sound, scheermes, uitsteken, learning, volgorde, vari, buurthuis	puppet, mmm, oppervlak, jacht, romein, koel, wrijven, warmer, period, uiterste

Table B.4: Words corresponding to 10 most distinguishing values in the PPMI vector of each control word in one community with respect to the PPMI vector in the other community

Bibliography

P. Cook, J. H. Lau, D. McCarthy, and T. Baldwin. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland, Aug. 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1154.

W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim. Bertje: A dutch bert model. *ArXiv*, abs/1912.09582, 2019.

M. Del Tredici and R. Fernández. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*, 2017. URL https://aclanthology.org/W17-6804.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

H. Dubossarsky, D. Weinshall, and E. Grossman. 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*, pages 1136–1145, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1118. URL https://aclanthology.org/D17-1118.

H. Dubossarsky, S. Hengchen, N. Tahmasebi, and D. Schlechtweg. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1044. URL https://aclanthology.org/P19-1044.

K. Ethayarajh. 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)*, pages 55–65, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1006. URL https://aclanthology.org/D19-1006.

J. Firth. A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*. Philological Society, Oxford, 1957. reprinted in Palmer, F. (ed. 1968) Selected Papers of J. R. Firth, Longman, Harlow.

L. Fremann and M. Lapata. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45, 2016. doi: 10.1162/tacl_a_00081. URL https://aclanthology.org/Q16-1003.

D. Geeraerts. *Diachronic Prototype Semantics: A Contribution to Historical Lexicology.* Oxford University Press, 1997.

D. Geeraerts, S. Grondelaers, and P. Bakema. *The Structure of Lexical Variation: Meaning, Naming, and Context.* Cognitive linguistics research. M. de Gruyter, 1994. ISBN 9783110143874. URL https://books.google.nl/books?id=8rwx7i3KZk8C.

D. Geeraerts, G. Kristiansen, and Y. Peirsman, editors. *Advances in Cognitive Sociolinguistics.* De Gruyter Mouton, 2010. ISBN 9783110226461. doi: doi:10.1515/9783110226461. URL https://doi.org/10.1515/9783110226461.

M. Giulianelli, M. Del Tredici, and R. Fernández. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.365. URL https://aclanthology.org/2020.acl-main.365.

D. Glynn. Semantics of sociolinguistic variation. a quantitative study of dialect effects on the polysemy of annoy. *Review of Cognitive Linguistics*, 8, 2010. ISSN 1877-9751.

K. Gulordava and M. Baroni. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK, July 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-2508.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. 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*, pages 2116–2121, Austin, Texas, Nov. 2016a. Association for Computational Linguistics. doi: 10.18653/v1/D16-1229. URL https://aclanthology.org/D16-1229.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, Aug. 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1141. URL https://aclanthology.org/P16-1141.

Z. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954. doi: 10.1007/978-94-009-8467-7_1. URL https://link.springer.com/chapter/10.1007/978-94-009-8467-7_1.

E. Harteveld. Fragmented foes: Affective polarization in the multiparty context of the netherlands. *Electoral Studies*, 71:102332, 2021. ISSN 0261-3794. doi: https://doi.org/10.1016/j.electstud.2021.102332. URL https://www.sciencedirect.com/science/article/pii/S0261379421000524.

R. Hasan. *Semantic Variation: Meaning in Society and in Sociolinguistics*. The collected works of Ruqaiya Hasan. Equinox, 2009. ISBN 9781904768364. URL https://books.google.nl/books?id=v3ScmgEACAAJ.

R. Hu, S. Li, and S. Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1379. URL https://aclanthology.org/P19-1379.

D. Jurafksy and J. H. Martin. *Speech and Language Processing (3rd ed. draft)*. 2021. URL https://web.stanford.edu/~jurafsky/slp3/.

Y. Kim, Y.-I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2517. URL https://aclanthology.org/W14-2517.

V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee. ISBN 9781450334693. doi: 10.1145/2736277.2741627. URL https://doi.org/10.1145/2736277.2741627.

A. Kutuzov and M. Giulianelli. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online), Dec. 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.14. URL https://aclanthology.org/2020.semeval-1.14.

W. Labov. *The Social Stratification of English in New York City*. Books on demand. Center for Applied Linguistics, 1966. ISBN 9780872811492. URL https://books.google.nl/books?id=NbpZAAAAMAAJ.

G. Lakoff. *The ALL NEW Don't Think of an Elephant!: Know Your Values and Frame the Debate*. Chelsea Green Publishing, 2014. ISBN 9781603585941. URL https://books.google.nl/books?id=aWhpBAAAQBAJ.

G. Lakoff, S. Farrar, and Giroux. *Whose Freedom?: The Battle Over America's Most Important Idea*. Farrar, Straus and Giroux, 2006. ISBN 9780374158286. URL https://books.google.nl/books?id=tOU7CWXqGDUC.

O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015. doi: 10.1162/tacl_a_00134. URL https://aclanthology.org/Q15-1016.

L. Lucy and D. Bamman. Characterizing English Variation across Social Media Communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556, 05 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00383. URL https://doi.org/10.1162/tacl_a_00383.

M. Martinc, P. Kralj Novak, and S. Pollak. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.592.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper.pdf.

O. Melamud, J. Goldberger, and I. Dagan. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany, Aug. 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1006. URL https://aclanthology.org/K16-1006.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

S. Mitra, R. Mitra, S. K. Maity, M. Riedl, C. Biemann, P. Goyal, and A. Mukherjee. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798, 2015. doi: 10.1017/S135132491500011X.

NCTV. Terrorist threat assessment for the netherlands, 2021. URL https://english.nctv.nl/documents/publications/2021/04/26/terrorist-threat-assessment-for-the-netherlands-54.

B. Noble, A. Sayeed, R. Fernández, and S. Larsson. Semantic shift in social networks. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 26–37, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.starsem-1.3. URL https://aclanthology.org/2021.starsem-1.3.

B. Obama. President obama's farewell address, 2017. URL https://obamawhitehouse.archives.gov/farewell.

E. Pariser. *The filter bubble : what the Internet is hiding from you.* Penguin Press, New York, 2011. ISBN 9781594203008 1594203008.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

J. A. Robinson. *Awesome insights into semantic variation*, pages 85–110. De Gruyter Mouton, 2010. doi: doi:10.1515/9783110226461.85. URL https://doi.org/10.1515/9783110226461.85.

J. A. Robinson. A gay paper: why should sociolinguistics bother with semantics?: Can sociolinguistic methods shed light on semantic variation and change in reference to the adjective gay? *English Today*, 28(4):38–54, 2012. doi: 10.1017/S0266078412000399.

M. Sahlgren and A. Lenci. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1099. URL https://aclanthology.org/D16-1099.

D. Schlechtweg, A. Hätty, M. Del Tredici, and S. Schulte im Walde. 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*, pages 732–746, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1072. URL https://aclanthology.org/P19-1072.

D. Schlechtweg, B. McGillivray, S. Hengchen, H. Dubossarsky, and N. Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online), 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.1. URL https://aclanthology.org/2020.semeval-1.1.

P. Shoemark, F. F. Liza, D. Nguyen, S. Hale, and B. McGillivray. 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)*, pages 66–76, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1007. URL https://aclanthology.org/D19-1007.

P. Sommerauer and A. Fokkens. Conceptual change and distributional semantic models: an exploratory study on pitfalls and possibilities. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 223–233, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4728. URL https://aclanthology.org/W19-4728.

C. Sunstein. *Echo Chambers: Bush V. Gore, Impeachment, and Beyond*. Princeton University Press, 2001. ISBN 9781400809059. URL https://books.google.nl/books?id=sEgHAAAACAAJ.

N. N. Tahmasebi. *Models and Algorithms for Automatic Detection of Language Evolution*. PhD thesis, Gottfried Wilhelm Leibniz Universität Hannover, november 2013. URL http://edok01.tib.uni-hannover.de/edoks/e01dh13/771705034.pdf.

E. C. Traugott. Semantic change, 03 2017. URL https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-323.

S. Ullmann. *Semantics : an introduction to the science of meaning / by Stephen Ullmann*. Blackwell Oxford, 1962.

L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

L. Wittgenstein. *Philosophical Investigations*. Basil Blackwell, Oxford, 1953. ISBN 0631119000.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Z. Wu, C. Li, Z. Zhao, F. Wu, and Q. Mei. Identify shifts of word semantics through bayesian surprise. In *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, SIGIR '18, page 825–834, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356572. doi: 10.1145/3209978.3210040. URL `https://doi.org/10.1145/3209978.3210040`.