Thema:

Computational Analysis of Semantic Shifts in Terms Reflecting Political Ideologies

Masterthesis

in the degree program Computing in the humanities offered by
the Faculty of Business Informatics and Applied Informatics at
the Otto-Friedrich University of Bamberg

Examiner: Prof. Dr. Roman Klinger

Name: Koch Fabian

Submission Date: 14.01.2026

Bamberg 2026

URN: urn:nbn:de:bvb:473-irb-114500x
DOI: https://doi.org/10.20378/irb-114500

Contents

List of Figures

List of Tables

1 Introduction and Motivation

Language is fundamental to human social life. It is not only a medium for transmitting information, but the primary means through which social reality is constituted and sustained. Shared meanings enable individuals to coordinate actions, deliberate collectively, and establish social and political institutions. Without a sufficient degree of semantic alignment, communication would lose its coherence and collective action would become fragile. At the same time, the question of how words acquire their meaning has long been a subject of debate across philosophy, linguistics, and cognitive science. Competing accounts locate meaning in reference to external entities, in mental representations, or in patterns of linguistic use within a community.

Despite these theoretical differences, there is broad agreement on one crucial point: meaning is not static. Word meanings evolve over time, shaped by historical developments, cultural change, and shifting social practices. As political systems transform and new challenges emerge, language adapts accordingly. Words accumulate new connotations, lose older associations, or are strategically reinterpreted, reflecting changes in collective values and worldviews.

Recent advances in computational linguistics have opened up new possibilities for studying such semantic change empirically. Within the framework of distributional semantics, meaning is operationalized through patterns of language use rather than through predefined symbolic definitions. One major part of distributional semantics and the methodical focus of this thesis are word embeddings, which are dense vector representations of meaning. These vectors provide a way to represent words in a continuous, low-dimensional space, where semantic relationships correspond to geometric relations. Through this, language becomes measurable data, and meaning can be analyzed as a structure within a mathematical space. The ability of word embeddings to capture the semantics and many other aspects of human language is widely acknowledged Kutuzov (2020).

This thesis tries to add onto these existing scientific endeavors and employs word embeddings to investigate semantic change in a domain that has received comparatively little attention in prior research: political speeches. Political language is especially well suited for such an analysis, as it is inherently normative and strategic. In politics, language does not merely describe reality; it actively frames problems, legitimizes actions, and mobilizes support. Words function simultaneously as instruments of communication and tools of persuasion.

Within this arena of discourse, one can observe an intriguing phenomenon: political actors often employ the same words to advance opposing visions of society. Terms such as "Freiheit" (freedom), "Frieden" (peace), and "Gerechtigkeit" (justivce) are invoked across the ideological spectrum, yet they hardly carry identical meanings as they are utilized to articulate fundamentally different ideas (McGee 1980). Their semantic nuances shift according to context, intent, and historical moment. By tracing these shifts computationally, this study seeks to uncover how the political use of language both reflects and reshapes the moral and ideological landscape of its user.

To this end, this thesis employs word embedding models to analyze political language across ideological contexts. Focusing on embeddings trained on ideology-specific corpora, the study examines how the semantic positioning of key political terms differs depending on the ideological context in which they are used. This thesis therefore sits in very close to a growing corpus of research called "diachronic word embedding" which focuses around the study of semantic change of words through a certain period of time.

Research Question

This study faces two main challenges: (1) detecting and quantifying semantic differences in word usage across political ideologies, and (2) interpreting these differences in a meaningful way. To address the first challenge, we develop three separate word embeddings for key political terms using three distinct text corpora, each representative of a specific political ideology: left-wing, right-wing, and center. These ideology-specific embeddings allow us to systematically analyze semantic shifts in word usage by employing the static embedding method Word2Vec for comparison. As training data, we use speeches delivered in the German Bundestag between 2017 and 2025-approximately 450 session transcripts containing 9,000 to 16,200 speeches. The data is publicly available, well-structured, and includes speaker affiliations, allowing us to assign ideological labels to the speeches based on party membership: right-wing (AfD), center (CDU/CSU), and left-wing (Grünen, Linke, SPD).

Differences in position alone provide little insight into how meanings change or in which direction they shift. Meaningful interpretation therefore requires reference concepts, so-called "anchor" points in the semantic space, that make it possible to assess whether terms move closer to or further away from specific semantic dimensions across ideological contexts.

In principle, such anchors can take various forms. One could, for instance, rely on other political concepts, policy-related terms, or manually curated ideological keywords to serve as reference points. However, these approaches often remain closely tied to the political domain itself and risk reproducing the same ambiguities that characterize political language. An alternative is to anchor semantic interpretation in more general, theoretically grounded constructs that are known to structure political attitudes but are not themselves narrowly political.

In this study, we adopt this latter strategy by drawing on Schwartz's theory of basic human values. The framework identifies ten universal values that have been empirically shown to structure human motivation and to correlate systematically with political orientation. For example, the value of "Macht" (power) tends to be more closely associated with conservative ideologies, whereas Universalism is more commonly linked to progressive ones. We hypothesize that these value-based associations are not only reflected in attitudes and beliefs, but also manifest in language use. Specifically, we expect the semantic positioning

of political terms to shift in ways that align with the dominant value systems of different ideological groups.

In this way, the present thesis moves beyond many existing applications of word embedding–based semantic change analysis, which often focus primarily on describing distributional shifts in language. While such studies can identify changes in semantic proximity or neighborhood structure of words, the interpretation of these patterns is frequently left implicit or tied closely to the linguistic domain itself. By grounding interpretations in a psychological theory of human values, this study takes an additional step and explicitly tries to connects patterns in language use to underlying motivational structures.Word embeddings are thus used not only to detect semantic variation, but to explore if or how psychological phenomenas can be mirrored in the use of political language. This leads us to the following hypothesis:

H: Our hypothesis is that the semantic meaning of key political terms is dependent on the political ideology for which they are invoked. We indicate this by showing that the semantic proximity between key political terms and values depends on whether the values correlate with the political ideology in whose context the key political terms appear.

Thesis Outline

This thesis is structured to gradually move from theoretical foundations to empirical analysis and interpretation. Following the introduction and motivation presented in Chapter 1, Chapter 2 provides a comprehensive overview of distributional semantics. It introduces the conceptual foundations of representing linguistic meaning computationally, tracing the development from symbolic and frequency-based representations to neural word embeddings. Special emphasis is placed on static word embedding models, which form the methodological core of this study, while contextualized embeddings are discussed to situate the chosen approach within the broader landscape of contemporary natural language processing.

Building on this theoretical background, Chapter 3 reviews existing research on semantic change within the framework of distributional semantics. It summarizes key findings from diachronic word embedding studies, discusses established regularities of semantic change, and examines ongoing debates about what embeddings capture when they are said to model "meaning." This chapter concludes by identifying a research gap, namely the limited attention given to ideological variation in political language within computational studies of semantic change.

Chapter 4 introduces Schwartz's theory of basic human values, which provides the conceptual foundation for interpreting semantic differences observed in the embedding spaces. The chapter explains why value theory is particularly relevant for political discourse and motivates its use as a psychologically grounded interpretative framework that goes beyond purely linguistic or domain-specific reference points.

Chapter 5 then details the methodological design of the study. It describes the data selection process, the construction of ideology-specific corpora based on speeches from the German Bundestag, and the training of word embedding models. The chapter also outlines the analytical procedures used to quantify semantic differences across ideological contexts.

The empirical findings are presented in Chapters 6 through 8. Chapter 6 reports results on the overall semantic stability and local semantic variation of political key terms. Chapter 7 introduces value embeddings and examines how value representations differ across ideological embedding spaces. Chapter 8 brings these strands together by analyzing the relationship between semantic change and value correlations using word-, sentence-, and axis-based approaches.

Finally, Chapter 9 discusses the results in light of the research questions and hypotheses. It reflects on the implications of the findings for understanding ideological meaning variation in political language, evaluates methodological choices, and situates the results within the broader literature. The thesis concludes by outlining limitations and directions for future research.

Results

The empirical analyses show that semantic variation across political ideologies is highly uneven and concentrates on specific types of political terms rather than affecting political language uniformly. Using stability measures and neighbor-based comparisons across ideology-specific word embedding spaces, the study finds that institutional and policy terms remain more semantically stable than cultural terms which terms exhibit a more pronounced ideological divergence. In particular, symbolic concepts related to identity and belonging like “Nation” or “Patriotismus” (patriotism) display substantial instability, indicating that ideological contestation operates primarily through the reinterpretation of culturally loaded vocabulary rather than through the redefinition of core institutional language.

This pattern is reinforced by the local neighbor-based analyses. The Jaccard-based neighbor divergence scores closely mirror the stability results: terms that are unstable in the global measure are also those whose nearest semantic neighborhoods diverge most strongly across ideologies e.g., “Leitkultur” (dominant culture) , “Nation” (nation), “Patriotismus” (patriotism). The qualitative neighborhood visualizations further illustrate that ideological differences often appear as shifts in framing and association, for example, where a term like “Asyl” (asylum) is embedded in humanitarian language in one corpus and more strongly linked to administrative control vocabulary in another.

The value-based analyses offer a more cautious picture. While term–value proximities vary across ideological contexts, these variations do not consistently align with the theoretically expected value–ideology associations derived from Schwartz’s framework. Instead of revealing stable ideological value poles in the embedding spaces, the results point

to fragmented and context-dependent value signals, particularly for contested cultural concepts. Taken together, the findings highlight both the strengths of word embeddings in detecting sites of ideological semantic contestation and the limitations of directly mapping psychological value theory onto distributional representations of political language. Future research may refine this connection by improving value operationalization or extending the analysis to other political genres and modeling approaches.

2 Computational Representations of Meaning

One of the central challenges in natural language processing (NLP) lies in bridging the gap between human understanding of language and computational representation. In other words, it concerns the transformation of linguistic symbols into numerical form. These numerical representations must preserve, as far as possible, the semantic and syntactic relations that exist between the original linguistic units.

Over the years, numerous techniques have been developed to achieve this transformation. The following section introduces the most influential approaches, presented in roughly chronological order, from early symbolic and frequency-based methods to modern neural architectures. Although this thesis employs a method situated between these two extremes, namely static word embeddings, a brief overview of earlier and later developments is provided for the sake of completeness and to highlight the conceptual progression of the field.

We begin with symbolic and frequency-based representations, including one-hot encoding and count-based models. Subsequently, we turn to neural approaches, with particular emphasis on static word embeddings like Word2Vec, which constitute the main focus of this study and will therefore be explained in detail. Finally, we conclude with a concise introduction to contextualized embeddings, which are generated through transformer models.

2.1 Symbolic Representations of Language

The earliest computational representations of language were symbolic in nature. In these systems, words were treated as discrete, atomic entities, unique symbols devoid of internal structure or semantic relation to other symbols. This approach reflected the broader tradition of symbolic artificial intelligence, in which knowledge and meaning were represented through explicit, human-defined rules and symbolic manipulation (Newell 1980). Within natural language processing (NLP), such representations took the form of one-hot vectors, where each word in the vocabulary is encoded as a binary vector of length V, with a single active element corresponding to the word's index and zeros elsewhere. Formally, this creates an orthogonal basis in a V-dimensional vector space, ensuring that any two distinct words are maximally dissimilar in representational terms.

Symbolic representations were foundational to early information retrieval and vector space models of text. The seminal work of Salton et al. (1975) introduced the Vector Space Model (VSM), which represented documents and terms as high-dimensional vectors and measured similarity through geometric operations such as the cosine distance. [1]

1 Whether Salton can be regarded as solely responsible for introducing the VSM is a matter of some debate. A more detailed discussion of this issue can be found in Dubin (2004).

This model established the mathematical groundwork for representing linguistic units in numerical form, allowing statistical operations to be performed on textual data. Subsequent refinements, including term frequency–inverse document frequency (TF–IDF) weighting (Salton and Buckley 1988), enhanced the discriminatory power of such representations by emphasizing words that are distinctive to particular documents.

Despite their practical utility, symbolic representations suffer from a fundamental limitation: they fail to capture semantic relatedness. Because each word is encoded independently, the vector for "king" bears no relation to that of "queen" or "monarch."

2.2 First Steps in Distributional Semantics: Frequency-Based Word Embeddings

This limitation of solely symbolic representations led to the development of distributional approaches, grounded in the assumption that the context of a word encodes its semantic information. This idea, formalized as the distributional hypothesis, holds that words occurring in similar contexts tend to have similar meanings. "You shall know a word by the company it keeps". This statement by John R. Firth still forms the bedrock of NLP technologies to this day (Firth 1957). Unlike symbolic systems, which treat words as isolated identifiers, distributional models try to capture meaning through statistical regularities in language use. By analyzing patterns of co-occurrence across large corpora, these models provide an empirical, data-driven foundation for studying semantics.

Early implementations represented words as frequency-based vectors derived from co-occurrence counts within a defined context window. Each dimension of such a vector corresponds to another word in the vocabulary, and its value reflects how frequently the two words co-occur.

Raw frequency counts, however, often overemphasize common but semantically uninformative words such as "the" or "is." To address this, weighting schemes like term frequency–inverse document frequency (TF–IDF) re-scale co-occurrence statistics by penalizing globally frequent words and emphasizing those that are more discriminative within particular contexts. Mathematically, for a word w in a document d,

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \log \quad \frac{N}{\text{DF}(w)}$$

where $TF(w, d)$ is the term frequency of w in d, N is the total number of documents, and $df(w)$ is the number of documents containing w. This weighting highlights words that are contextually specific, making the co-occurrence matrix more semantically informative. However, these raw matrices were typically large, sparse, and noisy, motivating the development of dimensionality-reduction techniques to extract latent semantic structure (Schütze 1992; 1998).

A key breakthrough came with Latent Semantic Analysis (LSA) (Deerwester et al. 1990), which applied singular value decomposition (SVD) to reduce co-occurrence matrices into lower-dimensional spaces. Given a term-document (or term-term) matrix X, LSA applies singular value decomposition:

$$X = U \quad V^\top$$

where U and V are orthogonal matrices representing word and document (or context) vectors, and is a diagonal matrix of singular values. By retaining only the top k singular values, we obtain a low-rank approximation

$$X_k = U_k \quad {}_k V_k^\top$$

which captures the most salient latent semantic dimensions. Words with similar distributions will then have similar representations in this reduced space. LSA demonstrated that statistical correlations in word usage could yield representations capturing semantic similarity and even analogy-like relations (Landauer and Dumais 1997). Subsequent models refined these methods by varying the way co-occurrence information was weighted and contextualized. Together, these approaches established the foundation of distributional semantics, in which meaning is emergent from linguistic patterns rather than predefined symbolic rules.

Frequency-based models not only provided a theoretical bridge between linguistics and computation but also marked a decisive shift in how meaning was conceptualized. They reframed semantics as a statistical phenomenon, paving the way for later neural embedding models that would learn such relationships directly from data. While limited by their reliance on global co-occurrence statistics and lack of context sensitivity, these early models represented a critical first step toward continuous, data-driven representations of meaning - a transition that culminated in neural word embeddings such as Word2Vec.

2.3 First Neural Distributional Semantics: Static Word Embeddings

While frequency-based distributional models demonstrated that statistical regularities in word co-occurrence encode meaning, their reliance on large, sparse matrices limited both efficiency and scalability. The next major development in distributional semantics was the emergence of neural embedding models, which sought to learn word representations directly from linguistic context rather than deriving them from pre-computed co-occurrence counts. These approaches produced dense, low-dimensional vectors that captured fine-grained semantic relations and offered dramatic improvements in both computational efficiency and representational quality.

The breakthrough came with the introduction of Word2Vec by Mikolov et al. (2013). In contrast to matrix-factorization methods such as LSA, Word2Vec employed a neural predictive model to infer vector representations that maximize the probability of observing context words given a target word (the Skip-gram model) or vice versa (the Continuous Bag-of-Words, CBOW). This training objective operationalizes the distributional hypothesis within a probabilistic framework: words that appear in similar contexts are adjusted toward similar representations in vector space. The resulting embeddings exhibit interpretable linear regularities, for instance, the vector relation "king – man + woman ≈ queen", suggesting that certain semantic and syntactic relations are encoded as vector offsets (Maslennikova and Bochkarev 2024).

Because Static Word embedding is the main method implemented in this thesis, we will explain the function of these models in detail. An understanding of Neural Networks and Backpropagation is necessary for the following. The main orientation for the following and the mathematical notations are derived from (Johnson et al. 2024), (Rong 2014) and of course (Mikolov et al. 2013).

2.3.1 Word2Vec: Architecture and Training Objectives

The basic framework of Word2Vec consists of a shallow, fully connected feed-forward neural network with one hidden layer. Mikolov et al. developed two main configurations with which the neural network can be trained: CBOW (Continuous Bag-of-Words) and Skip-gram. CBOW predicts a target word based on its surrounding context words and Skip-Gram conversely predicts the surrounding context words given a single target word. In both variants, each word w_i from a vocabulary V is represented initially as a one-hot vector $\mathbf{x}_i \in \mathbb{R}^{|V|}$, where only one component is 1 and all others are 0. Although both configurations share the same parameters namely two matrices:

$$W^{(1)} \in \mathbb{R}^{|V| \times N}, \quad W^{(2)} \in \mathbb{R}^{N \times |V|}.$$

where N is the embedding dimension (typically 100–300) they differ in the direction of what is predicted with these matrices and thus in what constitutes input and output during training.

2.3.2 Continuous Bag-of-Words (CBOW)

In the CBOW configuration, the model predicts a target word w_t from its surrounding context words $w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}$, where C is the context window size. As mentioned every one of these context words is initially represented as a one-hot vector. These vectors are projected into the embedding space through the input-hidden weight matrix $W^{(1)}$ and their embeddings are averaged to calculate a single hidden vector representation of the context. However for the sake of simplicity we will first assume that there

is only one word considered as context, which mean the model will predict one target word given one context word. Each row of $W^{(1)}$ is the N-dimensional vector representation v_w of the associated word of the input layer. Formally, row i of $W^{(1)}$ is v_w^T. Given a context (here as mentioned currently a single word) and assuming $x_k = 1$ and $x_{k'} = 0$ for $k^{'}$!= k, we have:

$$\mathbf{h} = W^{(1)\top}\mathbf{x} = W^{(1)\top}_{(k,\cdot)} := \mathbf{v}^{\top}_{w_I}$$

which is just copying the k-th row of $W^{(1)}$ to $\mathbf{h}$. $\mathbf{v}_{w_I}$ is the vector representation of the input word w_I. This implies that the link (activation) function of the hidden layer units is simply linear (i.e., directly passing its weighted sum of inputs to the next layer).

After we calculated the hidden state we use the weight matrix $W^{(2)}$ to compute a score for u_j for each word in the vocabulary. Remember that $W^{(2)}$ is a N x V matrix and the input is a vector with N-dimensions. As a result we get as many scores as words in V

$$u_j = W^{(2)}\mathbf{h} = \mathbf{v}'_{w_j}{}^{\top}\mathbf{h},$$

Here $\mathbf{v}'_{w_j}$ is the j-th column of the matrix W^2. We can then apply the softmax function, a log-linear model, to convert the scores into a probability distribution over all possible words in V:

$$p(w_j \mid w_I) = y_j = \frac{\exp(u_j)}{\sum_{j'=1}^{V}\exp(u_{j'})},$$

y_j here is the output of the j-unit in the output layer and represents the model's predicted probability that word w_j is the target word. During training, the model updates $W^{(1)}$ and $W^{(2)}$ to maximize the probability for the actual output word w_o '

$$\log p(w_o \mid \text{context}).$$

Hence, in theory CBOW performs a many-to-one prediction: several context words $\rightarrow$ one target word. Remember that until here we used a single context word to ease the introduction of the basic calculations.

To summarize, we developed a probabilistic formulation that enables the model to estimate how likely each word is to appear in a given context. The next step is to define an objective function that allows the model to learn these probabilities effectively by maximizing the likelihood of the target word given its context across the corpus. In other words, training the model - i.e., refining the semantic representations of words - involves maximizing the probability of the observed context words over the entire dataset D. To show this, we now derive the weight update equations for the model and show how the weight matrices are adjusted to improve these probabilities.

For this we first derive the weight update equation for the CBOW model. Although this computation becomes impractical for large vocabularies (as discussed later), deriving

it explicitly provides valuable insight into how the model adjusts its weight matrices to improve predicted probabilities before applying optimizations such as negative sampling or hierarchical softmax. The training objective is to maximize the probability of observing the actual output word w_O given the input context word w_I with regard to the weights:

$$\max p(w_O \mid w_I) = \max y_{j^*} = \max \log y_{j^*} = u_{j^*} - \log \sum_{j'=1}^{V} \exp(u_{j'}) := -E,$$

The loss function is defined as $E = -\log p(w_O \mid w_I)$, where j^* denotes the index of the true output word in the output layer. This objective can be interpreted as a specific instance of the cross-entropy measure between two probability distributions.

The process to generate the update equation for the weights stored in $W^{(2)}$ follows three main steps: 1. Compute the error at the output layer. 2. Derive the gradient of the loss with respect to the weights. 3. Use this gradient to update the weights using stochastic gradient descent (SGD). The error quantifies how wrong the model is, the gradient shows how the weights influence that error, and the update rule adjusts the weights to reduce future errors.

We begin by taking the derivative of the loss function E with respect to the net input of the j-th output unit, u_j:

$$\frac{E}{u_j} = y_j - t_j := e_j$$

i.e., t_j equals 1 only when the j-th unit corresponds to the actual output word; otherwise, $t_j = 0$. This derivative simply represents the prediction error e_j at the output layer. To calculate the gradient for the hidden -> output weights we next take the derivative on w'_{ij}:

$$\frac{E}{w'_{ij}} = \frac{E}{u_j} \cdot \frac{u_j}{w'_{ij}} = e_j \, h_i.$$

Therefore, using stochastic gradient descent, we obtain the weight update equation for the hidden -> output weights by:

$$w'^{(\text{new})}_{ij} = w'^{(\text{old})}_{ij} - \quad e_j \, h_i,$$

or equivalently,

$$\mathbf{v}'^{(\text{new})}_{w_j} = \mathbf{v}'^{(\text{old})}_{w_j} - \quad e_j \, \mathbf{h}, \quad \text{for } j = 1, 2, \ldots, V.$$

Here, > 0 represents the learning rate, $e_j = y_j - t_j$ denotes the prediction error, and h_i is the activation of the i-th hidden unit. The vector $\mathbf{v}'_{w_j}$ corresponds to the output representation of the word w_j. Note that this update rule entails iterating over all words in the vocabulary to evaluate their predicted probabilities y_j and compare them with the corresponding target values t_j (either 0 or 1). When $y_j > t_j$ (an "overestimation"), a fraction of the hidden vector $\mathbf{h}$ (i.e., $\mathbf{v}_{w_I}$) is subtracted from $\mathbf{v}'_{w_j}$, thereby pushing $\mathbf{v}'_{w_j}$ farther away from $\mathbf{v}_{w_I}$. In contrast, when $y_j < t_j$ (an "underestimation", which occurs only when $t_j = 1$, i.e., $w_j = w_O$), the hidden vector $\mathbf{h}$ is added to $\mathbf{v}'_{w_O}$, moving it closer to $\mathbf{v}_{w_I}$. If y_j is approximately equal to t_j, only minor changes occur. It should be emphasized that $\mathbf{v}_w$ and $\mathbf{v}'_w$ correspond to two separate vector representations of the same word w, serving distinct roles within the model.

After we calculated the update equations for $W^{(2)}$ we can move on to $W^{(1)}$. This is the embedding layer and where the final embeddings of words is stored. We follow similar steps as before: 1. Compute the error of the output at the hidden layer. 2. Derive the gradient of the loss with respect to the weights. 3. Use this gradient to update the weights using stochastic gradient descent (SGD). So first we take the derivative of E on the output of the hidden layer:

$$\frac{E}{h_i} = \sum_{j=1}^{V} \frac{E}{u_j} \cdot \frac{u_j}{h_i} = \sum_{j=1}^{V} e_j \, w'_{ij} := EH_i.$$

h_i is the output of the i-th unit of the hidden layer; u_j is defined as previous, as the input of the $j - th$ unit in the output layer and e_j = y_j - t_j is the prediction error of the j-th word in the output layer. EH is an N-dimensional vector and is the sum of the output vectors of all the words in the vocabulary, weighted by their prediction error.

With this we can formulate the derivative of E on $W^{(1)}$ obtaining:

$$\frac{E}{w_{ki}} = \frac{E}{h_i} \cdot \frac{h_i}{w_{ki}} = E_{H_i} \cdot x_k.$$

which is equivalent to the tensor product of X and EH:

$$\frac{E}{W} = \mathbf{x} \otimes \mathbf{EH} = \mathbf{xEH}^\top.$$

which computes a VXN Matrix. Remember that X is a one-hot encoded vector so in fact only one row of of $\frac{E}{W}$ is non-zero, and the value of that row is $\mathbf{E}_H^\top$, an N-dimensional vector. Thus, the update equation for W is given by:

$$\mathbf{v}_{w_I}^{(\text{new})} = \mathbf{v}_{w_I}^{(\text{old})} - \quad \mathbf{EH}^\top,$$

where $\mathbf{v}_{w_I}$ is a row of W, representing the input vector of the current context word. It is the only row of W whose derivative is non-zero; all other rows remain unchanged during this iteration since their gradients are zero.

The vector EH is the sum of all output word vectors, each weighted by its prediction error ($e_j = y_j - t_j$). This means that during training, the input vector of a context word is adjusted by portions of all output vectors in the vocabulary. If the model overestimates the probability of a word ($y_j > t_j$), the input vector moves away from that word's output vector; if it underestimates ($y_j < t_j$), it moves closer; and if the prediction is accurate, the movement is minimal. Over many training iterations, these adjustments accumulate so that each word vector is gradually pulled back and forth by its co-occurring neighbors, as if they were connected by invisible springs. Eventually, the "forces" between them balance, and the positions of the word vectors stabilize to reflect the patterns of co-occurrence in the training corpus.

Until now we only explained how CBOW worked in a single word context. In reality most CBOW configured Word2Vec models use 4 or more context words to calculate the embedding of a target word. However the differences arent huge. When computing the hidden layer output, instead of directly copying the input vector of the input context word, the CBOW model takes the average of the vectors of the input context words and uses this product as the vector which goes from the hidden to the output layer:

$$\mathbf{h} = \frac{1}{C} W^\top (\mathbf{x}_1 + \mathbf{x}_2 + \cdots + \mathbf{x}_C) = \frac{1}{C} (\mathbf{v}_{w_1} + \mathbf{v}_{w_2} + \cdots + \mathbf{v}_{w_C})^\top.$$

C is the number of words in the context, $w_1, \ldots w_c$ are the words in the context and v_W is the input vector a word w. The loss function and the update equation stay the same expect that $\mathbf{h}$ is defined as above. The update equation also is similar expect that now we apply this equation for every word in the context:

$$\mathbf{v}_{w_{I,c}}^{(\text{new})} = \mathbf{v}_{w_{I,c}}^{(\text{old})} - \frac{1}{C} \quad EH^\top$$

2.3.2.1 Skip-Gram

As mentioned Mikolov et al. (2013) did not only introduce the CBOW architecture but also the Skip-Gram configuration. Whereas CBOW predicts the target word from the context words, Skip-Gram inverts this relationship and predicts the contexts words from the target word.

Because skip-Gram at first is just CBOW with a single word as context we can copy the definition of **h** from above. **h** is simply copying (and transposing) a row of the input -> hidden weight matrix, $W^{(1)}$, associated with the input word wI:

$$\mathbf{h} = W_{(k,\cdot)}^{(1)\top} := \mathbf{v}_{w_I}^{\top},$$

On the output layer however instead of outputting one multi nominal distribution, we are outputting as many multi nominal distributions as words we want to predict from our initial word w_I denoted by C. Each output is again computed by, using the hidden -> output matrix $W^{(2)}$:

$$p(w_{j,c} = w_{O,c} \mid w_I) = y_{c,j} = \frac{\exp(u_{c,j})}{\sum_{j'=1}^{V} \exp(u_{j'})}.$$

In this formulation, $w_{c,j}$ refers to the j-th word computed by the c-th output head of the model. The variable $w_{O,c}$ denotes the true c-th context word associated with the input word w_I. The output $y_{c,j}$ is the predicted probability assigned to the j-th vocabulary word by the c-th output head, and $u_{c,j}$ is the corresponding pre-softmax activation value.

Since all output heads share the same weight matrix, the activation for the j-th unit can be written as

$u_{c,j} = u_j = {\mathbf{v}'_{w_j}}^{\top}\mathbf{h}$, for $c = 1, 2, \ldots, C$.

Here, $\mathbf{v}'_{w_j}$ represents the output vector of the j-th word in the vocabulary w_j, which corresponds to the j-th column of the hidden-to-output weight matrix $W^{(2)}$.

With regard to the derivation of the parameter of the update equation we again can see the similarity between Skip-Gram and the One-Word-Context CBOW model:

$$E = -\log p(w_{O,1}, w_{O,2}, \ldots, w_{O,C} \mid w_I) = -\log \prod_{c=1}^{C} \frac{\exp(u_{c,j_c^*})}{\sum_{j'=1}^{V} \exp(u_{j'})} =$$

$$-\sum_{c=1}^{C} u_{j_c^*} + C \cdot \log \sum_{j'=1}^{V} \exp(u_{j'}).$$

Here j_c^* is the index of the actual c-th output context word in the vocabulary. We take the derivative of E with regard to the net input of every unit on every panel of the output layer u_c^j and obtain:

$$\frac{E}{u_{c,j}} = y_{c,j} - t_{c,j} := e_{c,j}$$

which is the prediction error on the unit same as in CBOW. We will define a V-dimensional $EI = EI_1, \ldots EI_V$ which represents the sum of prediction error over all context words:

$$E_{I_j} = \sum_{c=1}^{C} e_{c,j}$$

Now we can take the derivative of E with regard to the hidden->output matrix $W^{(2)}$:

$$\frac{E}{w'_{ij}} = \sum_{c=1}^{C} \frac{E}{u_{c,j}} \cdot \frac{u_{c,j}}{w'_{ij}} = E_{I_j} \cdot h_i$$

which is the same as:

$$\mathbf{v}'_{w_j}{}^{(\text{new})} = \mathbf{v}'_{w_j}{}^{(\text{old})} - \quad E_{I_j}\,\mathbf{h}, \qquad \text{for } j = 1, 2, \ldots, V.$$

One can understand this update equation similar to that of CBOW with the difference that the prediction error is summed across all context words in the output layer. Note that if we wouldn't use efficiency algorithms to increase the performance of the training, we would need to apply this update equation for every element of $M^{(2)}$ for every training instance. That means that in a vocabulary of 100,000 words, the original softmax formulation forces us to touch 100,000 output vectors per training sample, so a single epoch over a million-word corpus already requires roughly 10^{11} weight updates.

The update equation for $W^{(1)}$ is the same as in CBOW with the only difference that the singular prediction error e_j is replaced with EI_j.

$$\mathbf{v}_{w_I}^{(\text{new})} = \mathbf{v}_{w_I}^{(\text{old})} - \quad EH^\top$$

EH is an $N-$dim vector of which each component is defined as:

$$EH_i = \sum_{j=1}^{V} EI_j\, w'_{ij}$$

The interpretation of EH is here the same as in CBOW. EH is the sum of all output word vectors, each weighed by its prediciton error. Again, during training the input vector is adjusted by portions of all output vectors in V.

2.3.2.2 Computational Optimization in Word2Vec

In the above formulations of the bigram, CBOW, and Skip-gram models, no efficiency optimizations are applied. Each word has both an input vector and an output vector, and while updating the input vectors is computationally cheap, updating the output vectors is as mentioned extremely costly. According to the update equations, training requires iterating over the entire vocabulary for every training instance to compute each word's score, softmax probability, prediction error, and corresponding gradient update. This results in $O(|V|)$ operations per training example, which becomes impractical for large vocabularies or corpora. To address this bottleneck, techniques such as hierarchical softmax and negative sampling reduce the number of output vectors that must be updated per instance. These methods specifically optimize the computation of output-layer updates while preserving the required gradients for the objective function and backpropagation.

Hierarchical softmax (HS) replaces the flat softmax with a binary-tree factorization of the output distribution: predicting w_O becomes a sequence of binary decisions along the path from the root to the leaf corresponding to w_O. Only the parameters on this path are involved in computing the probability and gradients, substantially reducing computation compared to evaluating all V outputs. Conceptually, hierarchical softmax changes how the model computes the probability that each vocabulary word is the correct output word given the current input word (or context) but it still trains the same input and output representations through backpropagation from an output-layer objective.

Negative sampling (NS) avoids normalization over all of V by training the model to distinguish observed (positive) target–context pairs from a small number of sampled (negative) pairs. For each instance, parameters are updated only for the true word(s) and the sampled negatives, yielding a large efficiency gain while still pushing embeddings to reflect co-occurrence structure.

The full mathematical notation, derivations, and update formulas for both Hierarchical Softmax and Negative Sampling are provided in the appendix.

The previous section gave an in-depth explanation of Word2Vec. Despite its limitations by today's standards and the rapid progress in NLP since its introduction, Word2Vec remains a remarkable milestone in the development of distributional semantics. As explained, before its emergence, most computational approaches relied either on high-dimensional, sparse co-occurrence matrices or on symbolic representations that failed to capture meaningful similarities between words. Word2Vec fundamentally shifted this paradigm by introducing an efficient neural architecture capable of learning dense, low-dimensional

embeddings directly from local linguistic context. Its ability to encode analogical relationships and fine-grained semantic regularities with unprecedented computational efficiency represented a genuine breakthrough. Even though later models, especially transformer-based architectures, have surpassed it in representational power, Word2Vec paved the way for modern embedding methods and remains significant as the moment when neural distributional semantics became both practical and widely accessible. Word2Vec is still a relevant part of NLP research today and continues to serve as the methodological backbone of many scientific endeavors (Pak et al. 2024; Asudani et al. 2023; Yash Mahajan 2025).

2.4 Context is King: Contextualized Word Embeddings with Transformers

While static word embedding methods like Word2Vec offer a strong foundation, they also highlight the need for something more flexible. This is where transformers enter the picture. Transformers are widely seen as one of the biggest breakthroughs in modern NLP, and they motivate the shift toward contextualized word representations, where meaning is treated as dynamic rather than fixed.

2.4.1 The Attention Mechanism

Modern transformer-based architectures (Vaswani et al. 2017) changed the field by replacing recurrent processing with self-attention mechanisms. Instead of giving every word a single "static" vector, transformers compute a new embedding for each specific occurrence of a word dynamically from its current context. This allows models like BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) to pick up on subtle differences in meaning across sentences, speakers, or domains. Because transformers take the entire surrounding context into account instead of a predefined context window, they can handle things like polysemy, pragmatic cues, and complex syntactic patterns much more effectively than earlier methods. Their success has led to widespread use in tasks ranging from sentiment analysis and information extraction to historical semantics (Giulianelli et al. 2020).

This shift from static embeddings, as used in Word2Vec, to fully contextualized representations is achieved in transformers through a series of learned linear transformations applied to each token's embedding within a sentence: After the input tokens are mapped to their initial embedding vectors, conceptually similar to Word2Vec, each embedding is projected into three separate spaces, producing a query q, a key k, and a value v vector through multiplication with the corresponding weight matrices ($W_Q, W_K and W_V$). These results of the multiplication of the embedding with the corresponding weight matrix encode different aspects of the token's representation depending on its role in the attention

computation. The queries derived through W_Q encode what information a token seeks, the keys produced via W_K specify what information each token provides to others, and the values obtained from WV contain the information that can be integrated when attention is applied.

For example, take the word "at" in the sentence "She arrived at the station." After "at" is turned into its initial embedding, each weight matrix shapes that embedding in a different way. The projection with the query matrix W_Q can highlight what "at" is looking for in the sentence, for instance, that it usually needs a location or noun to complete its meaning. The projection of the key matrix W_K can highlight what "at" offers to other words, such as signaling that it is a preposition connecting two parts of the sentence. The value matrix W_V produces a version of "at" that carries the information the model should use if this word becomes important during attention. In summary, even a small word like "at" is turned into three different representations that help the model understand how it relates to the surrounding words.

These projected vectors are collected into matrices Q, K, and V and passed into the central formula of "Attention Is All You Need":

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

This computes a weighted combination of the value vectors, where each weight reflects how compatible a query is with the corresponding key. The result is a matrix whose rows represent new, context-dependent vectors for each token. In practice, transformers compute several such attention functions in parallel, a procedure known as multi-head attention, allowing the model to capture different types of relationships simultaneously (e.g., syntactic, semantic, or discourse-related patterns).

After attention is computed, each token's updated vector is passed through a small feed-forward neural network applied independently to every token. This simply provides another learned transformation and does not require sequential structure. Transformers also use residual connections, meaning that the attention output (after its linear transformation) is added back to the original input embedding. This helps preserve the token's initial semantic information while enriching it with context-sensitive cues derived from the attention mechanism.

The resulting vectors, which now combine both a token's initial representation and its interactions with surrounding tokens, form the contextualized embeddings that are passed to the next layer of the transformer.

2.4.2 Bigger is Better - The Emergence of LLMs

The development of contextualized embeddings through attention not only transformed how dynamically words and sentences can be modeled, but also set the stage for a new generation of language models: pretrained language models. BERT was among the first to show how powerful this idea can be in practice (Devlin et al. 2019). Its architecture consists of multiple transformer encoder layers (i.e., layers that read an input sequence and iteratively build richer representations of each token without generating new text) that repeatedly refine each token's embedding by integrating information from all other tokens in the sentence. During pre-training, BERT uses a bidirectional masking objective: selected tokens are replaced by a mask symbol, and the model must reconstruct them using both left and right context. This forces the embeddings to encode meaning based on the entire linguistic environment rather than a fixed window or direction. This design allowed BERT to achieve high performance on a wide range of downstream tasks and helped establish the "pre-training and fine-tuning" paradigm that remains central to NLP today (Zhao et al. 2023).

BERT's success inspired a large number of follow-up studies, many of which explored variations of the underlying transformer architecture. One prominent example is GPT-2, which replaces BERT's encoder-based design with a decoder-only architecture (i.e., a stack of transformer layers tailored for autoregressive next-token prediction rather than bidirectional context integration). This shift places the focus on generating text token by token, allowing the model to learn rich contextual embeddings through prediction-driven training-instead of relying on BERT's bidirectional masking objective (Radford et al. 2019). Another major line of research focused not only on architectural changes but also on improving pre-training strategies (Sanh et al. 2021; Wang et al. 2022). RoBERTa (Liu et al. 2019) is a prominent example: it retains BERT's encoder-based architecture but optimizes almost every aspect of its training procedure. Instead of introducing a new model structure, RoBERTa removes BERT's next-sentence prediction objective, trains on significantly larger corpora, uses much larger batch sizes, and applies dynamic masking rather than fixed masking patterns. These adjustments do not change how embeddings are computed within the transformer layers, but they substantially improve the quality and stability of the resulting contextualized word representations. It did not take long for the NLP community to recognize that scaling model size, data, and compute was one of the most effective ways to improve performance, ultimately leading to the formulation of empirical scaling laws (Kaplan et al. 2020). These laws show that a language model's test loss decreases in a predictable, smooth power-law manner as parameters, training data, and compute increase, meaning that larger models consistently perform better, provided they are trained with proportionally sufficient data and computational resources.

This rapid growth in model size ultimately gave rise to the term large language models, with the release of ChatGPT marking one of the most influential moments in bringing advanced AI capabilities into broad public use (Ouyang et al. 2022; Shanahan 2024). This transition did not merely reflect an increase in parameter count, it marked a fundamental

shift in how language models were understood as tools. Earlier statistical language models primarily served narrow, well-defined purposes such as information retrieval or speech recognition, where probability estimates improved specific downstream pipelines. The first neural language models then expanded this scope by learning task-agnostic linguistic features, reducing the need for handcrafted features but still functioning largely as components within specialized systems.

The introduction of pretrained transformer-based models extended this paradigm further. Models like BERT and RoBERTa learned deep, contextualized representations that could be efficiently adapted to a wide range of downstream tasks through fine-tuning, positioning pretrained LMs as flexible feature extractors rather than single-purpose tools. With the emergence of LLMs, however, the role of language models changed again: scaling model size, data, and compute yielded systems capable not only of representing language but of solving diverse, open-ended tasks far beyond traditional NLP benchmarks. These models are increasingly viewed as general-purpose problem solvers - systems that can reason, summarize, generate, translate, and interact conversationally - rather than as mere generators of text or encoders of linguistic structure.

In short, the historical trajectory from statistical LMs to neural LMs, to pretrained LMs, and finally to LLMs reflects a steady broadening of task scope and a dramatic increase in model capacity. Each stage extended what language models could do: from predicting words, to extracting features, to adapting across tasks, to performing complex, multi-step reasoning and providing human-facing assistance. This conceptual shift from modeling language to solving tasks captures the central evolution in the research history of language models.

However, since universal task-solving LLMs exceed the scope of the present study, we end our theoretical chapter here. In the following section, we turn to concrete applications of diachronic word embeddings and outline the current state of research in this field.

3 Distributional Semantics in Action: State of the Field

Given the centrality of semantics to language, communication, and human interaction more broadly, it follows that a wide range of research communities engage with distributional semantics, most notably computational linguistics, information retrieval, and humanities disciplines such as political science (Kutuzov et al. 2018).

It is therefore unsurprising that research on how word meanings evolve predates contemporary computational approaches by many decades (Bréal 1899; Stern 1931). Among the early systematic treatments, Bloomfield's work (Bloomfield 1933) occupies a central position. Rather than treating semantic change as a single phenomenon, he proposed a taxonomy of nine recurrent patterns, six of which form three complementary pairs. The first pair-*narrowing* and *broadening* (widening)-captures shifts in the scope of a word's reference. Broadening occurs when a term becomes more general, as with English "dog", which developed from Middle English "dogge", originally referring to a specific breed. Conversely, narrowing can restrict a formerly general term to a more specific sense, illustrated by Old English "mete" (food), which later specialized to Modern English "meat" (edible flesh).

A second set of opposing tendencies involves change in expressive force: *hyperbole*, where a term strengthens in intensity, and *meiosis*, where it weakens. While Bloomfield illustrated these processes with various historical examples, the mechanisms themselves have since been widely recognized in later semantic-development studies. A third evaluative dimension contrasts *elevation* with *degeneration*. Elevation involves the acquisition of more positive or prestigious associations, exemplified by the development of "knight" from Old English "cniht" (boy, servant); degeneration refers to the drift toward more negative connotations.

Bloomfield also identified three non-oppositional mechanisms that reflect shifts in how speakers conceptualize referents: *metaphor*, the transfer of meaning based on perceived similarity; *metonymy*, meaning shifts based on contiguity or association; and *synecdoche*, where a part–whole (or whole–part) relation motivates semantic extension. Although these categories have since been refined and expanded, they remain foundational. Subsequent work emphasizes that such semantic shifts are shaped by an interplay of communicative pressures, cognitive tendencies, and broader sociocultural developments (Blank and Koch 1999; Grzega and Schoener 2007).

Broadly linguists separate semantic shifts into two important classes: linguistic drift (gradual, systematic changes in a word's core meaning that stem from the internal dynamics of the language) and socio-cultural shifts (changes in meaning shaped by evolving social practices, cultural norms, and speakers' shifting associations with a term). Linguistic shifts are semantic changes that arise from pressures internal to the language system itself - such as phonological erosion, grammatical reanalysis, analogical extension, or semantic bleaching - leading to slow but regular adjustments in a word's conventional sense over time (Sapir 1921). Socio-cultural semantic shifts, by contrast, are changes in word meaning which are driven by non-linguistic social or cultural factors, for example, technological developments (Hamilton et al. 2016). With regard to this thesis, the focus is on socio-cultural

shifts, since these changes allow scholars to draw broader conclusions about underlying cultural phenomena.

Whereas the above considerations were primarily philosophical and of a qualitative nature, the introduction of computational linguistics transformed the landscape and enabled more systematic, quantitative data-driven approaches. It soon became evident that computational approaches offered a powerful means to empirically test long-standing hypotheses about the law like nature of semantic change (Nölle et al. 2020). In the following section, we present key research in this area.

3.1 Laws of Semantic Change

One intuitive example is the research of Dubossarsk et al. (2015). They examine whether a word's position within its semantic category, specifically, whether it is a more central or more peripheral member, can predict how likely it is to undergo semantic change. Using a bottom-up approach, they train yearly word embeddings on English fiction data from 1850 to 2009 and apply K-means clustering, a method that groups words into clusters based on their similarity by assigning each word to the nearest cluster center. Each cluster has a centroid, the mathematical average of all vectors in that group, which serves as an abstract prototype for the category. For every word, the authors then measure it's *prototypicality* as its distance to this centroid and quantify semantic change through self-similarity across consecutive decades. Their key finding is a strong negative correlation: words closer to their category's centroid change less over time, while those farther away display greater semantic drift. In later research papers they refer to this phenomena as the "*law of prototypicality*," suggesting that semantic stability is shaped by a word's relative centrality within its category (Dubossarsky et al. 2017).

Another interesting research is from Xu and Kemp (2015), who conducted a computational evaluation of two classic but competing laws of semantic change: *the law of differentiation* (Bréal 1899), which proposes that near-synonyms tend to diverge in meaning over time and *the law of parallel change*, which claims that words with related meanings tend to change in similar ways over time (Lehrer 1985). For this they constructed word embeddings through co-occurrence patterns in the Google N-gram corpus from 1890 to 1999, representing each word by a normalized distribution over 5,000 context words using Jensen–Shannon divergence as a measure of semantic similarity. Semantic change was quantified by comparing a word's nearest neighbors across decades, and synonym and antonym pairs from historical and modern thesauri were evaluated against carefully matched control pairs. Their results show that semantically related words overwhelmingly tend to move in parallel rather than diverge, providing strong computational support for the *law of parallel* change while offering little evidence for the *law of differentiation*.

Building on this growing body of work, Hamilton et al. (2016) took a broader and more systematic approach to identifying regularities in semantic evolution (Hamilton et al. 2018). Instead of examining *prototypicality* or *synonym-based dynamics* as in the previous

studies, they focus on two broader predictors of semantic change - word frequency and polysemy - and ask whether these factors give rise to general statistical laws that apply across languages. To explore this, they construct diachronic word embeddings for six historical corpora across four languages, using several embedding methods (PPMI, SVD, and SGNS). Semantic change is then quantified as the cosine distance between a word's embeddings in consecutive time periods, and these change rates are regressed against their linguistic predictors. Their analysis reveals two robust cross-linguistic trends. First, *the law of conformity*, according to which frequent words tend to change more slowly. Second, *the law of innovation*, which shows that once frequency is controlled for polysemous words change significantly faster than monosemous ones. In contrast to earlier case-study-based work, their findings demonstrate that these effects scale consistently across languages and centuries, providing strong quantitative evidence that both frequency and polysemy systematically shape the trajectory of semantic change.

A more critical perspective is offered by Dubossarsky et al. (2017), who argue that several of the proposed "laws" of semantic change may be largely artefacts of the distributional methods used to detect them (Dubossarsky et al. 2017). Their central methodological claim is that any genuine law of semantic change must appear in real historical data but disappear in a corpus where no semantic change is possible. To test this, they introduce two control conditions: a chronologically shuffled corpus, in which all text snippets are redistributed across decades so that each decade contains an identical mix of contexts, and a synchronous corpus, created by repeatedly sampling from the same year. In both controls, true semantic change is impossible, meaning that any observed "change" must stem from statistical noise or biases in the representation model. When re-evaluating the previously proposed laws, *the law of conformity* (frequent words change less), *the law of innovation* (polysemous words change more), and *the law of prototypicality* (prototypical words change less), they find that the same correlations persist almost as strongly in the control conditions as in the genuine historical corpus. This demonstrates that these effects arise not from actual semantic evolution but from artefacts introduced by count-based vector representations, cosine similarity, and frequency-related noise. As a result, they conclude that the strong versions of these laws do not withstand rigorous validation and call for more stringent methodological standards in the study of semantic change.

3.2 Diachronic Word Embedding in Action

While the previous section focused on theoretical attempts to identify general laws that govern how words change meaning over time, the following section turns to concrete applications of these ideas. Instead of asking whether semantic change follows universal statistical principles, we now examine how diachronic word embeddings can be used practically to trace meaning shifts in specific domains, corpora, and historical contexts. As mentioned earlier, diachronic word embeddings are simply word embeddings trained on successive time slices of a corpus so that each time slice captures a distinct "snapshot" of semantic structure. By aligning these embeddings across time, researchers can directly

observe how the neighborhoods of words expand, contract, or reorganize, thereby revealing semantic shifts in a quantifiable way. Following the categorization of Kutuzov et al. (2018) there are broadly two possibilities for this kind of approach: *linguistic studies* which investigate the how and why of semantic shifts, and *event detection* approaches which mine text data for actionable purpose. The first category generally involves corpora with longer time spans, since linguistic changes happen at a relatively slow pace whereas the second category involves mining texts for cultural semantic shifts (usually on shorter time spans) indicating real-world events.

An illustrative example for a linguistic study is the work by Hamilton et al. (2016), who use diachronic word embeddings to trace how the sentiment of words shifts across both historical periods and social communities. They introduce *SentProp*, a method that combines distributional embeddings with label propagation to induce sentiment lexicons from unlabeled corpora using only small sets of seed words. Applying this framework, they construct community-specific sentiment lexicons for 250 major Reddit communities and historical sentiment lexicons spanning 150 years of English. Their analysis reveals striking patterns of semantic change: over 5 percent of sentiment-bearing words completely reversed polarity over the last one and a half centuries - for example, “terrific” shifted from strongly negative to strongly positive (*elevation*) - while others, such as “lean” and “pathetic”, underwent clear amelioration and pejoration (*degeneration*).

Another early application of diachronic analysis is the study by Mihalcea and Nastase (2012), who introduce word epoch disambiguation: the task of predicting the historical period in which a word occurrence appears. Using Google Books data, they extract contextual snippets for 165 target words across three fixed epochs (around 1800, 1900, and 2000) and train a Naive Bayes classifier using only local and topical contextual features. The aim is to test whether changes in usage stemming from shifts in sense, vocabulary, or topics are strong enough to allow automatic classification of the words to the epochs. Their model clearly outperforms their baseline (61 percent vs. 43 percent accuracy), and performance is highest for polysemous words and words with strong frequency differences across epochs, indicating that both contextual and sense-related changes leave detectable traces over time (Mihalcea and Nastase 2012).

Another notable contribution is from Jones et al., who used diachronic embeddings to investigate gendered associations of words related to career, family, science, and art. Their study, based on word embeddings trained on two centuries of books, found that many traditional gender stereotypes in these domains have diminished in language usage over time (Jones et al. 2020).

For event detection research a simple and intuitive example is the study of Braun who tracked semantic shifts in German Court Decisions. He applied diachronic word embeddings to analyze how legal language changes over time, especially in response to legislative changes (Braun 2022). Using a corpus of over 200,000 German court rulings from 1970 to 2020, he trained 59 different word embeddings using Word2Vec across various time periods. His findings revealed that legislative changes (events) can cause almost instantaneous semantic shifts in court language. For example, he documented how the

meaning of "Lebenspartner" (life partner) shifted after the Civil Partnership Act of 2001. Before 2000, the most similar word was "Lebensgefährten" (a synonym without legal implications). After 2000, the most similar word became "Ehepartner" (spouse), indicating a shift toward a legally recognized relationship. He observed similar shifts in the embeddings of the word "Wehrdienst"(military service) or "Rundfunkbeitrag" (broadcasting fee).

A similar study is from Kulkarni et al. (2015) who trained word embeddings for successive time periods across three large datasets (Twitter posts, Amazon product reviews and Google Books) and aligned these embeddings into a unified space to track how each word's position changes over time and thereby uncovering well-known historical changes, such as the shift of "gay" from meaning cheerful to homosexual, and detected rapid, domain-specific changes like "streaming," or "combo" acquiring new senses in response to technological innovations or real-world events (Kulkarni et al. 2015).

A particularly relevant study for our topic is that of Hellrich et al. (2018), who analyzed the temporal dynamics of emotions associated with specific words. They showed that beginning in the 1990s, the word "climate" began to evoke increasingly negative emotions, while the emotional intensity associated with it also rose over time (Hellrich et al. 2018).

3.3 Meaning in Vector Space: What do Word Embeddings Represent?

A final group of studies we want to introduce takes a more meta-theoretical perspective on word embeddings. Rather than applying diachronic models to specific linguistic or cultural questions, these papers ask what word embeddings actually represent and what kinds of meaning they allow us to study. Their focus is not on practical NLP applications but on the conceptual foundations of distributional semantics - examining how language is encoded in vector spaces, which aspects of meaning are captured, and which are systematically distorted or lost. Such work provides an essential framework for understanding the methodological assumptions underlying diachronic word embeddings and for clarifying what kinds of semantic change they can actually validly measure.

Central to this line of inquiry is the scope and implications of the distributional hypothesis. As noted earlier, the idea that "you shall know a word by the company it keeps" (Firth 1957) underlies all computational approaches in NLP. Yet there remains debate about how deeply this principle should be taken. If one assumes that all semantic meaning is fully encoded in the contexts in which a word appears - no matter how broadly one defines "context" - then, in principle, computational models might be capable of capturing the very essence of language. Such a view implies that words acquire their meanings solely through recurring patterns of co-occurrence, and that two words occurring in similar linguistic environments *thereby* share a similar meaning (Arseniev-Koehler and Foster 2022). This raises profound questions: Is meaning nothing more than statistical regularity, or is there something irreducible that escapes contextual capture? The way we answer this shapes

not only what we can expect from our models, but how we conceptualize language itself. As Boutyline and Arseniev-Koehler (2025) state, only a minority of semanticists subscribe to such a strong form of the distributional hypothesis [2]. With this in mind, we must ask the question what we actually capture with our embeddings and how does this differ from what we want to measure. To answer these questions we will roughly follow the paper of Boutyline & Arseniev-Koehler (2025).

3.3.1 What do we Measure?

Word embeddings do not exclusively capture semantic meaning; they reflect whatever regularities are present in the contexts from which they are derived. As Boutyline & Arseniev-Koehler (2025) emphasize, distributional models encode patterns of usage, and these patterns often arise from surface-level linguistic structure rather than semantic content. Embeddings therefore pick up syntactic and morphological regularities as well as nonsemantic co-occurrences including orthographic variation, stable phrases, stylistic conventions, and even artifacts like OCR (Optical Character Recognition) errors or typographical habits. Words may appear close in a vector space not because they share conceptual properties, but because they repeatedly occur together in formulaic expressions ("French fries") or genre-specific contexts (e.g., months clustering with cities due to news article opening format e.g., "CHICAGO, June 10, 2019). This means that embeddings represent contextual associations, only some of which correspond to semantic meaning. For many practical NLP applications, these unintended associations pose less of a problem than they do for sociological or cultural research. After all, such "false" connections - where words co-occur without sharing any underlying semantic property - are not merely artifacts of computational models but are also reflected in human associative behavior. As Nelson et al. (2004) demonstrate, when people are asked to freely associate words, they are more likely to pair "French" with "fries" than with "people," even though only the latter two relationships reflect conceptual similarity. So for a next-token-predicition model it is actually good when "french" lies closely to "fries" as there is a certain probability that this is the token we want. Nonetheless, we can see how co-occurrence–based associations can arise from habitual language without necessarily being semantically meaningful. The challenge, then, is to disentangle meaningful relations - those that reflect real semantic connections - from coincidental or artifactual co-occurrences. Understanding this distinction is crucial, because many embedding-based studies ultimately aim to measure not mere statistical proximity but the conceptual meaning of social categories, practices, or objects. This leads us to our second question.

2 For a more detailed discussion of the meaning and the limits of the distributional hypothesis, see Sahlgren (2008).

3.3.2 What do we want to Measure: Context vs. Concept

Boutyline and Arseniev-Koehler's (2025) systematic review of sociological embeddings-based research shows that the answer to the question above varies considerably across studies. Of the 39 empirical articles they review, about a quarter use embeddings simply as measures of linguistic regularity rather than meaning. In this work, embeddings quantify similarity across texts or capture stylistic and formal novelty and their sensitivity to non-semantic patterns is a feature rather than a flaw.

The remaining three-quarters of studies, however, rely on embeddings as measures of semantic meaning - for instance, to estimate the cultural associations of wealth or studiousness, the stigma of illnesses, the religiosity of names, or occupational prestige. Just as the research we presented above these applications require that distances in embedding space reflect semantic similarity rather than mere co-occurrence, making it essential to distinguish genuinely semantic information from the many nonsemantic regularities present in context spaces.

To make this distinction possible we need to clarify what kind of semantic structure embedding-based research aims to recover. Drawing on the theoretical models of so called *concept spaces*, Boutyline and Arseniev-Koehler argue that what researchers ultimately seek to measure is not the embedding's context space itself, but an underlying concept space: an abstract, multidimensional structure whose dimensions correspond to the socially meaningful features people use to understand, categorize, and evaluate concepts (Gärdenfors 2004; 2011). This is the core distinction between context and concept spaces: the interpretability of their dimensions. Unlike context spaces, whose dimensions are arbitrary artifacts of model training, concept spaces encode interpretable properties such as size, status, morality, gender, or symbolic value. A context space is therefore a valid estimator of conceptual meaning only to the extent that its geometric structure is isomorphic to the structure of the relevant concept space, that is, only insofar as relative distances between words in the embedding reflect relative proximities between the corresponding concepts.

This theoretical framework has two major implications. First, because embeddings represent words as single points while concepts correspond to fuzzy, multidimensional regions, embeddings necessarily provide a simplified approximation of meaning: at best, they capture the prototypical core of a concept rather than its full variability. Second, concepts differ in which dimensions are even relevant to them. Many concepts have values on only a limited subset of feature dimensions, and concepts belonging to different domains may not be comparable along the same axes. These properties complicate the task of using context spaces to measure meaning but also underscore the importance of conceptual clarity when interpreting embedding-based results.

Taken together, these observations suggest a simple but important lesson for NLP: embeddings cannot be treated as ready-made maps of meaning. Because they reflect all kinds of contextual regularities: semantic and nonsemantic alike. Therefore researchers

Type of space	What it is	Entities within the space	Meaning of coordinates
Context space	Abstract many-dimensional space that summarizes the tendency of terms to be used together with other terms; the space that word embeddings directly estimate.	Terms from the corpus (usually individual words or frequent phrases; sometimes subword units, stems, etc.).	Coordinates are not directly meaningful; two terms are proximate if they occur in similar usage contexts.
Concept space	A abstract many-dimensional space that represents concepts in terms of their semantically meaningful features (e.g., size, status, gender).	Abstract, language-independent concepts.	Coordinates correspond to the features people use to comprehend, categorize, and evaluate concepts.

Table 3.1: Comparison of context spaces and concept spaces. Adapted from Boutyline & Arseniev-Koehler (2025).

who use them to study cultural meaning must actively separate the two. Boutyline and Arseniev-Koehler emphasize that doing so requires clearer theorizing about what we want embeddings to measure, more careful operationalizations that target concepts rather than surface word forms, and systematic validation against human judgments. In their view, embeddings can provide powerful insights into meaning, but only when researchers treat them as noisy, context-based estimates that must be refined rather than as direct windows onto conceptual structure.

3.4 Research Gap: Bridging Distributional Semantics and Psychological Value Theory

From Diachronic Meaning Change to Synchronous Semantic Variations

The preceding chapter outlined key developments in diachronic word embedding research, linguistic theories of semantic change, and recent meta-theoretical debates concerning what embeddings represent and what kinds of meaning they can reliably measure. With this in mind and the research laid out, we can formulate the way how this thesis addresses key research gaps and tries to fill them.

One of the most striking features of the current state of the field is that with few exceptions like Azarbonyad et al. (2017) it almost exclusive focuses on diachronic semantic change. Nearly all computational research examines how meaning changes across historical time whether its concerned with laws of semantic evolution, lexical drift, or cultural shifts (Mihalcea and Nastase 2012; Jones et al. 2020). Studies compare embeddings trained on the 1850s with those of the 1950s; track how "gay" shifted from "cheerful" to "homosexual" (Kulkarni et al. 2015); or measure how sentiment attached to concepts such as "climate" or "technology" evolved over decades (Hellrich et al. 2018). Even when research adopts more fine-grained temporal slices, such as Twitter data covering only a few years, the logic remains rooted in temporal comparison: meaning is conceptualized as something that changes across time (Guo et al. 2021).

This diachronic orientation leaves a considerable blind spot. It overlooks the possibility that semantic variation may also occur synchronically, across different segments of a society at a single historical moment. Political actors, social groups, ideological communities, or subcultures often use the same words while implicitly referring to different concepts. "Freedom" may signify individual autonomy for one group and collective self-determination for another; "Justice" may evoke distributive fairness for some and meritocratic equality for others. These divergences are not temporal in nature at least not primarily but ideological. Existing embedding-based research, however, provides no systematic method for detecting such synchronous conceptual divergence, even though it is a central premise in political psychology, political theory, and discourse studies.

This thesis therefore shifts the analytical lens from diachronic change to synchronous ideological variation, investigating whether political groups in the German Bundestag employ core political concepts in meaningfully different ways. By constructing ideology-specific embedding spaces from parliamentary speeches and comparing their semantic structures, the thesis examines how meaning varies not across time, but across political worldviews.

Lack of Embedding-Based Research on Political Conceptual Meaning

A second research gap concerns the limited application of embedding-based methods within political science. While NLP has been used extensively for political tasks such as classifying ideology (Iyyer et al. 2014), measuring sentiment in political statements (Mohammad et al. 2017), predicting vote intention (Tumasjan et al. 2010), or detecting hate speech (Davidson et al. 2017), these studies rarely investigate what political concepts mean in context. Instead, embeddings are used instrumentally, as features in statistical classifiers or as tools for document similarity.

There is almost no research that places the meanings of political concepts at the center of analysis. Political communication scholarship regularly notes that parties "frame" issues differently, highlight distinct aspects of a concept, or appeal to different moral values

(Entman 1993). Yet these claims are rarely operationalized using modern semantic methods. Schwartz's value theory demonstrates that value priorities differ systematically across the political spectrum. Since values guide how people understand and evaluate social realities, these differences imply that political concepts like "freedom" or "justice" may take on distinct value-laden meanings for different ideological groups. This line of research, however, has not been tested or explored systematically with word embeddings.

Existing studies lack the research for estimating how political concepts - such as "Macht" (power), "Freiheit" (freedom), or "Solidarität" (solidarity) are embedded within the ideological meaning systems of political actors. No study to date examines whether these concepts occupy different semantic neighborhoods when used by conservative, moderate, or progressive political speakers within the same institutional setting.

This thesis fills this gap by reframing embeddings as tools for conceptual political analysis. It is not concerned with predicting ideology from language, but with understanding how ideology shapes the meaning of language itself. Through a comparative embedding framework, it evaluates whether political concepts drift toward ideologically consistent value dimensions, thereby offering a computational grounding for long-standing claims in political theory and psychology.

Absence of a Value-Theoretic Framework in Semantic Embedding Studies

A third gap concerns the lack of a value-theoretic interpretive framework in embedding research. Even when semantic differences are detected - whether diachronic or cross-domain - the interpretation of these differences often remains ad hoc. Researchers typically rely on nearest-neighbor lists, sentiment lexicons, or informal inspection of cosine distances. While such methods can identify surface-level associations, they rarely address deeper conceptual questions. What dimensions of meaning differentiate concepts across groups? Do these differences correspond to underlying psychological values, cultural schemas, or political ideologies?

Schwartz's value theory for example provides a well-established psychological framework for understanding the motivational underpinnings of political attitudes. Yet such theories have not been integrated into embedding-based research on semantic meaning. This omission is striking: value theory is fundamentally concerned with the conceptual organization of human cognition and thus offers a natural theoretical lens for interpreting how ideological groups imbue political concepts with meaning.

Embeddings as Imperfect Proxies for Conceptual Meaning: The Need for Theoretical Clarity

The fourth major gap arises not from political science per se but from the broader theoretical debate about what embeddings actually represent. As discussed in the previous chapter, word embeddings capture context spaces, not concept spaces. They reflect statistical distributional patterns that mix semantic, syntactic, stylistic, and nonsemantic regularities. While recent sociological work has emphasized this distinction, few empirical studies operationalize it explicitly.

This creates a methodological tension: researchers often use embeddings as if they directly encode conceptual meaning, while in reality, embeddings only approximate meaning to the extent that their geometric structure is isomorphic to the underlying conceptual space. For politically loaded concepts, this distinction is especially important. A word such as "Demokratie" (democracy) may appear near similar neighbors across ideological corpora not because its conceptual meaning is stable, but because shared parliamentary procedural vocabulary masks underlying ideological divergences.

This thesis contributes to resolving this issue by trying to apply the concept–context distinction from Boutyline and Arseniev-Koehler (2025) to political semantics. A more detailed explanation of this follows in the methods chapter. The project carefully separates surface-level contextual proximity from concept-level meaning, using Schwartzs value dimensions to assess whether embedding distances reflect interpretable ideological associations rather than mere lexical co-occurrence. In doing so, the thesis offers both a methodological clarification and a practical demonstration of how embeddings can be aligned with conceptual rather than purely contextual meaning.

Limited Embedding-Based Research on German Parliamentary Discourse

Finally, the existing literature shows a pronounced linguistic and geographic bias. Most embedding-based work relies on English-language corpora: Google Books, COHA, Reddit, Twitter, or U.S. congressional speech. Only a small number of studies examine non-English political discourse, and virtually none employ embeddings to investigate political semantics in German.

This gap is particularly relevant given the importance of parliamentary speeches in shaping German political discourse and the rich ideological diversity represented in the Bundestag. German political concepts - such as "Freiheit" (freedom), "Solidarität" (solidarity), or "Ordnung" (order) - carry historical, cultural, and ideological connotations that may diverge substantially across parties.

This thesis fills this gap by constructing , for the first time, a computational map of how political concepts are positioned within the ideological meaning structures of German

political language. It provides a foundation for future research in German political semantics and contributes to the comparative study of ideological meaning across linguistic contexts.

Taken together, these gaps show that while semantic research has advanced our understanding of meaning over time, it has largely overlooked how meaning varies across political ideologies in the present. This thesis addresses this oversight in four key ways. First, it introduces an approach for analyzing synchronous ideological semantic variation, examining how meanings differ across political ideologies rather than across historical periods. Second, it uses embedding models to approximate the conceptual meaning of key political terms, rather than merely measuring their contextual co-occurrence. Third, it incorporates Schwartz's value theory trying to build interpretable semantic dimensions that ground these differences in established psychological structures. Finally, it provides the first embedding-based study of German parliamentary discourse, mapping how German political actors conceptualize central political terms.

Collectively, these contributions link computational linguistics with political psychology and political communication/ideology research. They show that embedding models - when embedded in a clear conceptual framework - can reveal how political actors inhabit distinct semantic worlds even while using the same vocabulary, thereby advancing our understanding of political meaning both methodologically and substantively.

4 Shalom Schwartz's Value Theory

To address the research gaps identified above, particularly the need to interpret ideological differences in political meaning, this thesis requires a theoretical framework that specifies which latent dimensions structure these semantic variations. Schwartz's theory of basic human values provides such a framework. The following chapter introduces this value theory in broad detail and is structured into four parts: first, an overview of Schwartz's model, secondly, the relationship between values and political orientation, third, reasons for preferring Schwartz's framework over alternative value theories, and lastly the implications for embedding-based semantic analysis.

4.1 Why Values Matter for the Semantics of Political Debates

Political debates are rarely disagreements about facts alone; more often, they reflect deeper differences in how individuals and political communities understand and evaluate the world. Core political concept - as freedom, justice or solidarity - are not simply descriptive terms but carry a normative weight. The semantic of these terms may in part reflect the underlying values that political actors seek to advance. Two groups may use the same vocabulary yet refer to fundamentally different conceptual structures, grounded in distinct value orientations.

Understanding such differences is essential for interpreting political language. A liberal and a conservative speaker may invoke the term "freedom", but what each understands by this concept differs considerably. One may emphasize autonomy, self-expression, or non-interference, while the other associates it with order, responsibility, or moral restraint. These conceptual divergences cannot be explained by linguistic structure alone; they emerge from deeper value systems, which function as interpretive lenses that shape how individuals perceive and categorize political ideas. This thesis tries to encapture these values from the language of the political speaker.

Because of this, value theory plays an important role. When we examine whether different ideological groups in the German Bundestag assign distinct meanings to political concepts, we require a theoretical framework that specifies how meaning varies. What underlying dimensions might structure these differences? Schwartz's theory of basic human values provides such a framework. It offers a well-established, empirically grounded account of the motivational domains that orient human judgment, behavior, and political orientation. Moreover, its structural and dimensional nature makes it uniquely suitable for embedding-based semantic analysis. Before turning to the empirical methods of this thesis, it is therefore necessary to introduce Schwartz's value theory, outline its core claims, and explain why it offers an appropriate conceptual foundation for studying ideological semantic variation.

4.2 Overview of Schwartz's Theory of Basic Human Values

Schwartz's value theory, first formulated in 1992 and refined in subsequent decades (Schwartz 1992; Baudisch 2018), emerges from the attempt to identify universal values that guide human decision-making across different cultures. With values as a core element in understanding human behavior and opinion formation, Schwartz builds on a tradition shaped by foundational social theorists such as Durkheim and Weber (Schwartz 2012b). Schwartz characterizes values with six core features they all share: 1. Values are beliefs that are linked inextricably to affect and that become infused with feeling when activated. 2. Values refer to desirable goals that motivate action. 3. Values transcend specific actions and situations. This distinguishes values from narrower concepts like norms and attitudes. 4. Values serve as standards or criteria. They guide the selection or evaluation of actions, people, policies, and events. People judge what is good or bad, justified or illegitimate, based on its possible consequences for their cherished values. This usually occurs outside conscious awareness. 5. Values are ordered by relative importance, forming a system of priorities that characterizes each individual. This hierarchical feature also distinguishes values from norms and attitudes. 6. The relative importance of multiple values guides action. Any attitude, opinion or behavior typically has implications for more than one value. The trade-off among relevant, competing values is what guides attitudes and behavior (Schwartz 1992).

The Ten Basic Values according to Schwartz

Taken together, these six features define the general properties that all values share. Values express fundamental psychological needs either biological, interpersonal, or societal and help individuals navigate trade-offs between competing motivations. What distinguishes one value from another is the specific type of goal or motivation it represents. Based on this principle, Schwartz formulated ten distinct values:

1. Power – social status, dominance over people and resources.
2. Achievement – personal success through demonstrating competence.
3. Hedonism – pleasure and sensuous gratification.
4. Stimulation – excitement, novelty, and challenge.
5. Self-Direction – independent thought and action; creativity, autonomy.
6. Universalism – understanding, tolerance, and protection of all people and nature.
7. Benevolence – concern for the welfare of close others.
8. Tradition – respect and commitment to cultural or religious customs.
9. Conformity – restraint of actions that might upset others or violate norms.

10. Security – safety, stability of society, interpersonal relationships, and self.

These values are not random categories but reflect universal human concerns that vary in priority across individuals and cultures. However, these values are not independent of one another but can be organized along two higher-order dimensions that structure their interrelations. "Self-Enhancement" <-> "Self-Transcendence" (power and achievement contrast with universalism and benevolence) and "Openness to Change" <-> "Conservation" (self-direction and stimulation contrast with security, conformity, and tradition). Values associated with "Openness to Change" promote autonomy in thinking, feeling, and acting, as well as a willingness to embrace new experiences. In contrast, "Conservation" values emphasize self-restraint, adherence to long-standing traditions, and a preference for order and stability. "Self-Transcendence" values focus on recognizing the equality of others and showing concern for their well-being, whereas "Self-Enhancement" values center on personal achievement, advancement, and asserting influence over others. Hedonism occupies a position between "Self-Enhancement" and "Openness to Change", sharing features of both but generally aligning more strongly with the latter (Schwartz 2012a). With this ordering of the values one can create the following motivational cirlce:

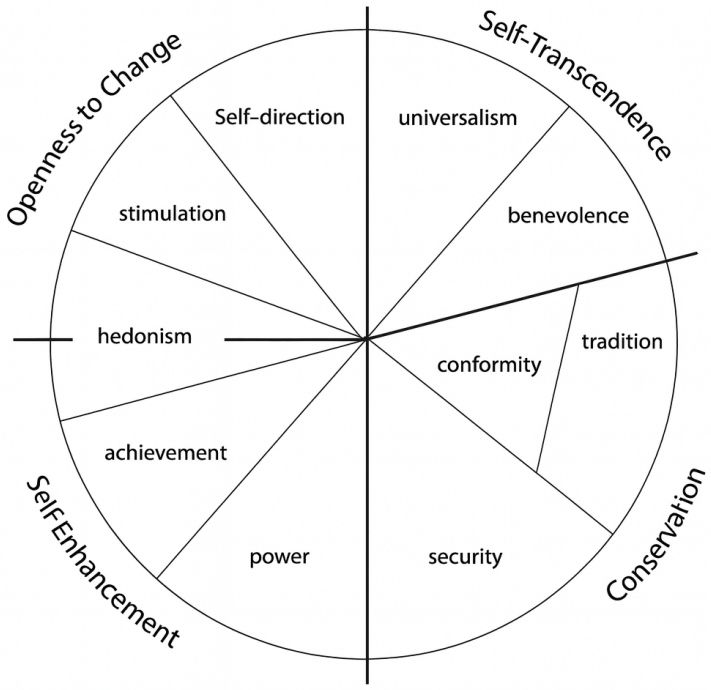

Figure 4.1: Schwartz's circumplex model of basic human values (Schwartz 1992).

This structure implies that values in adjacent positions are motivationally compatible, while those across from each other express conflicting motivations. For example, pursuing power often conflicts with universalism, and prioritizing tradition may conflict with stimulation.

4.3 Empirical Research: The Correlation between Values and Political Orientations

One of the reasons Schwartz's theory gained such broad acceptance is its empirically demonstrated stability. Studies across more than 80 cultural contexts consistently reproduce the circumplex structure (Boer and Fischer 2013). Though cultures differ in the emphasis placed on specific values, the relational configuration of values remains remarkably constant.

Given this stability, it is unsurprising that researchers have also examined how value priorities relate to political orientation. If values shape people's interpretations of political concepts, then ideological differences should correspond to systematic differences in value profiles. Decades of empirical research support this claim: the association between Schwartz's values and political ideology has been replicated extensively across national contexts, demographic groups, and political systems.

Across these studies, a consistent pattern emerges. Individuals with right-wing or conservative orientations tend to place greater emphasis on "Self-Enhancement" values such as power and achievement, as well as "Conservation" values including security, tradition, and conformity. By contrast, those with left-wing or progressive orientations generally assign higher priority to "Self-Transcendence" values like universalism and benevolence, along with "Openness to Change" values such as self-direction and stimulation.

Right-wing values	**Left-wing values**
Power	Universalism
Achievement	Benevolence
Security	Self-Direction
Tradition	Stimulation
Conformity	

Table 4.1: Idealized Associations between political ideologies and value priorities based on Schwartz's value theory.

Which specific values become most salient in ideological discourse often depends on the political issues that dominate a given context. In the Israeli political context of the late 1980s for example, where debates centered on the balance between religious tradition and secular freedoms, the principal value differences among party supporters lay in their endorsement of tradition versus the values self-direction and universalism (Barnea and Schwartz 1998). A similar pattern appeared in the Italian national elections of 2001, though the relevant values shifted because the political issues were different. Here, the center-right emphasized entrepreneurship, market liberalism, security, and traditional family and national values, whereas the center-left foregrounded social welfare, social justice, equality, and tolerance toward diverse groups. Correspondingly, the values most

strongly associated with political preference differed from those observed in Israel: security and power were more prominent among right-leaning voters, while universalism and benevolence were most characteristic of those supporting the center-left. Values accounted for impressive 18% (Caprara et al. 2006) of the total variance in political preference, whereas age, income, education and gender together only contributed 2%. Lastly, in a study of political preferences in 14 different countries, Barnea (2003) found that where political competition revolved around issues of national security vs. equal rights, the key values that structured voters' preferences were security and conformity vs. universalism and self-direction. However, where political competition revolved around the distribution of material resources, the key values were universalism and benevolence vs. power and achievement (Barnea 2003).

The relationship between values and political ideology is well established at the level of attitudes and preferences, but it can be expected that the connection between values and political ideology extends beyond mere attitudes and voting decisions and reaches even into the realm of conceptual meaning. The underlying hypothesis of this thesis is that individuals' value priorities may subtly guide how they interpret and evaluate political concepts; meaning that their values shape which aspects of a concept they emphasize, which associations they consider natural or legitimate, and which trade-offs they regard as morally acceptable. In this view, political vocabulary does not have a fixed meaning but is filtered through value-laden interpretive frames.

For instance, individuals who place strong importance on Security and Tradition might be inclined to interpret Freedom as freedom from disorder or external threat, whereas those who prioritize Self-Direction may understand Freedom primarily as the ability to pursue personal goals and express oneself autonomously. This hypothesized link between value priorities and conceptual meaning provides the motivation for the methodological approach adopted in this thesis: if values inform how political actors understand key political terms, then differences in value orientations should leave detectable traces in embedding-based semantic representations. Of course the approach taken in this thesis departs from Schwartz's original psychological framework, as it seeks to capture values on a linguistic level by representing them as embeddings and testing whether the value–ideology patterns documented in prior research can be replicated within the semantic vector space.

4.3.1 The Suitability of Schwartz's Value Theory for Embedding-Based Semantic Analysis

Several competing frameworks could potentially serve as theoretical foundations for studying political meaning, including Inglehart's postmaterialism index (Inglehart 2015), Rokeach's value survey (Rokeach 1973), or Haidt's Moral Foundations Theory (Haidt and Graham 2007). Yet Schwartz's value theory is uniquely appropriate for integrating with embedding-based semantic methods.

Conceptual Dimensionality and Geometric Structure

Schwartz's circumplex provides continuous, interpretable dimensions, which align naturally with geometric properties of vector spaces. Values such as power, universalism, or security can be represented as points or axes in embedding space, allowing political concepts to be projected along meaningful psychological dimensions. Other frameworks (e.g., Rokeach) lack such a dimensional structure and instead produce lists of values without defined relational geometry.

Explicit focus on conceptual meaning instead of concrete judgments

While other frameworks such as Haidt's Moral Foundations Theory also involve conceptual categories, their focus is primarily on explaining moral intuitions - rapid affective judgments about right and wrong - rather than the broader conceptual structures that organize political meaning. Schwartz's theory, in contrast, defines values as cognitive representations of abstract goals that guide how people interpret and evaluate a wide range of social and political concepts. Because these values function as general conceptual dimensions rather than domain-specific moral reactions, they align more naturally with the representational structure of word embeddings, which capture broad patterns of semantic association across contexts.

Cross-Cultural Validity and strong empirical Link to Ideology

Because Schwartz's value structure has been validated across a wide range of cultural contexts, it provides a stable foundation for conceptual dimensions relevant to German political discourse. His values exhibit robust and replicable correlations with political orientations across countries and over time, making them well suited for capturing systematic differences in how political communities conceptualize core terms.

Operational Flexibility of Schwartz's Values in Embedding Space

Perhaps most importantly for this thesis, Schwartz values can be represented in multiple embedding formats: Word-based representations e.g. "Macht" (power), "Universalismus" (universalism), Sentence-based embeddings using textual descriptions and Axis-based embeddings, using antonymic contrasts (e.g., Power – Weakness) to construct conceptual dimensions. More details on the embedding methods in the following chapter.

For these reasons, Schwartz's model provides a suitable theoretically grounded and methodologically compatible framework for analyzing ideological semantic variation

using word embeddings. Integrating his theory into embedding-based semantic analysis provides a crucial interpretive layer for understanding ideological meaning differences. Value dimensions help distinguish genuine conceptual divergence from mere co-occurrence patterns and allow embedding distances to be linked to underlying psychological orientations rather than remaining opaque geometric artifacts. By grounding the analysis in a well-established value framework, this thesis can interpret semantic variation in a way that connects linguistic usage to deeper psychological needs, revealing how different ideological groups may ascribe distinct meanings to the same political terms.

5 Research Methodology

Having established the theoretical foundations, the state of research in distributional semantics, and Schwartz's value theory, we now turn to the methodological framework for our empirical analysis. The preceding chapters have shown, first, that word meanings are not fixed but depend on the linguistic and social contexts in which they are used; and second, that political orientations correlate systematically with distinct value priorities. The central task of this thesis is to bring these two insights together: to examine whether ideological differences in political discourse lead to systematically different semantic realizations of key political concepts, and whether these differences align with the value structures that political psychology has documented extensively.

Our core hypothesis builds on this synthesis. If political actors prioritize different values such as power, security, universalism, or benevolence then these priorities may influence how they conceptualize central political terms. We therefore expect that the semantic of terms like ”Freiheit” (freedom), “Gerechtigkeit” (justice)), or “Solidarität” (solidarity) will shift depending on whether they appear in conservative, moderate, or progressive discourse. Concretely, the hypothesis states that the semantic proximity between political terms and Schwartz's value concepts will vary systematically across ideological contexts: political terms used in conservative speeches should cluster more closely with values associated with conservative orientations (such as Power or Security), whereas the same terms used in progressive speeches should align more strongly with values such as Universalism or Benevolence.

Operationalizing this hypothesis requires a methodological strategy that captures ideological variation in language at the level of distributional semantics and connects it to the value-based conceptual dimensions described by Schwartz. To do so, we construct separate embedding spaces for three ideological contexts - conservative, moderate, and progressive - based on speeches from the German Bundestag. These embeddings serve as models of how each political camp employs and interprets key political concepts. We then derive linguistic representations of Schwartz's values through several embedding strategies and measure how political terms converge toward these value vectors across ideological contexts.

The remainder of this chapter details this methodological approach step by step. We first describe the operationalization of our main concepts - political key terms and value dimensions - before explaining how we construct and compare ideology-specific embedding models. We then discuss decisions regarding training data, embedding methods, and evaluation procedures.

5.1 Operationalization of Political Concepts

Ideological Viewpoints: Right-Wing, Centrist, and Left-Wing

To examine whether political ideology shapes the semantic structure of political language, we must first define how both ideologies and concepts are represented in our empirical

framework. Following Azarbonyad et al. (2017), who conceptualize a viewpoint as a corpus segment that shares a particular metadata feature, such as political party, time period, or social group, we treat ideology as a linguistic viewpoint that can be operationalized through separate, ideologically segmented corpora. In their study, semantic differences emerge when embeddings are trained on speeches from distinct political parties, demonstrating that Conservative and Labour MPs attribute systematically different meanings to key political terms (Azarbonyad et al. 2017). We will orientate ourselves by their approach and construct ideological viewpoint by partitioning the speeches held in the German Bundestag into three distinct groups: right-wing, centrist and left-wing. This classification is based on party affiliation in the German Bundestag. Each ideological group is treated as a distinct viewpoint and is used to train its own embedding space. The resulting embeddings should therefore reflect not only the linguistic patterns shared within each ideological group but also the conceptual emphases and associations characteristic of that group.

Our decision to partition the corpus along a single political spectrum into right-wing, centrist, and left-wing categories is motivated by multiple factors. Although political belief systems can indeed be multidimensional with for example economic, social, and foreign policy dimensions often treated as analytically distinct (Carsey and Layman 2006) a substantial body of research shows that the left–right axis remains the dominant latent structure of political conflict and party competition in Western democracies. Work in political psychology demonstrates that citizens and elites tend to integrate diverse issue positions into a broadly coherent left–right worldview (Jost et al. 2008), and studies in political behavior find that left–right self-placement reliably predicts positions across both economic and cultural domains (Knight 2006). Even studies that explicitly adopt multidimensional frameworks, such as Sinno et al. (2022), acknowledge that left–right remains the most stable and interpretable organizing dimension, with additional dimensions emerging primarily in fine-grained analyses of specific policy issues rather than in general political discourse (Sinno et al. 2022a).

From a methodological perspective, the use of a single left–right partition also aligns with the requirements of distributional semantic modeling. Training separate embedding spaces for multiple ideological subgroups demands sufficiently large and internally coherent corpora. As Azarbonyad et al. (2017) shows, fragmenting the data into too many ideological categories reduces corpus size and therefore model stability, making semantic comparisons unreliable. Parliamentary speech, moreover, rarely articulates political ideology in neatly separable economic or cultural dimensions: arguments about migration, national identity, social justice, or economic fairness frequently blend multiple value-laden domains within the same rhetorical context. A three-way division into right-wing, centrist, and left-wing blocs therefore captures the major ideological cleavages present in parliamentary communication while preserving the statistical robustness needed for training high-quality embedding models.

Selection of Key Political Terms

The next step is to define which political terms will be analyzed across viewpoints. Because our theoretical interest lies in how ideological orientations shape the meaning of core political vocabulary, we focus on three categories of terms central to democratic discourse: institutional terms, policy terms, and cultural terms.

Institutional terms denote the foundational structures of the political system. Although often treated as stable or "neutral," political theory shows that such concepts are essentially contested (Gallie 1955; Freeden 1996). Different ideological actors emphasize different aspects of institutional legitimacy, authority, or responsibility, producing divergent semantic meanings even for terms that refer to shared democratic institutions. As institutional terms we analyse: *Demokratie (democracy), Freiheit (freedom), Grundgesetz (basic law), Verfassung (constitution), Bundestag (Bundestag), Bundesregierung (federal government), Opposition (opposition), Sozialstaat (welfare state), Föderalismus (federalism), Rechtsstaat (rule of law)*

Policy terms represent domains of political action with high ideological salience. Because these issues involve normative trade-offs and activate competing moral and interpretive frames, they are particularly likely to reveal clear semantic divergence between right-wing, centrist, and left-wing viewpoints (Lakoff 2022). The meaning of such terms is inherently shaped by broader value commitments and partisan priorities. As policy terms we analyse: *Integration (integration), Migration (migration), Asyl (asylum), Umweltschutz (environmental protection), Klimaschutz (climate protection), Energiewende (energy transition), Frieden (peace), Sicherheit (security), Inflation (inflation), Gerechtigkeit (justice)*

Cultural terms are especially prone to semantic contestation because they carry symbolic, affective, and identity-laden connotations. Prior research demonstrates that these concepts are among the most ideologically malleable, reflecting enduring struggles over national identity, collective belonging, and the boundaries of political community (Brubaker 2020; Giesen and Seyfert 1999). Their meaning is often central to broader ideological narratives about society and nationhood. For the cultural terms we decided for: *Würde (human dignity), Solidarität (solidarity), Heimat (homeland), Vaterland (fatherland), Nation (nation), Volk (people), Identität (identity), Leitkultur (guiding culture), Patriotismus (patriotism), Europa (Europe), Deutschland (Germany)*

Selecting these terms allows us to probe a wide spectrum of political meaning from institutional concepts to normative policy issues and deeply symbolic cultural notions.

5.2 Operationalization of Political Terms and Values

Having outlined the political key terms whose semantic variation we want to examine, we now turn to the question of how these terms and the values of Schwartz are operationalized within oour embedding space. In practice, this means extracting or constructing vector

representations from our Word2Vec models, which then serve as the empirical basis of the analysis. While political concepts can be represented directly through their word embeddings, for example, by using the vector associated with "Deutschland" (Germany), Schwartz's values pose a greater challenge, as they refer to abstract motivational goals rather than concrete lexical items. To capture these value dimensions in language, we therefore rely on three complementary strategies: (1) simply using the embeddings of the value labels themselves e.g., the vector for "Tradition" (tradition), (2) constructing sentence-level embeddings from textual definitions of the values, and (3) deriving semantic axes from pairs of contrastive anchor terms. Each of these approaches captures a different aspect of value meaning, and together they provide a more robust operationalization than any single method alone.

Word based Embedding

The simplest operationalization treats each value (e.g., "Macht" (power) , "Universalismus" (universalism), "Tradition" (tradition) as a lexical item and extracts its vector directly from the embedding space. This representation captures how the value term is used in parliamentary discourse itself. Its main advantage is conceptual transparency: if political actors, for example, use the term "Gerechtigkeit" (justice) in systematically different contexts across ideological groups, these differences will be reflected in their respective embeddings. More concretely, "Gerechtigkeit" (justice) may appear closer to "Sicherheit" (security) in one ideological discourse suggesting a semantic emphasis on order, stability, or protection while in another it may be embedded farther away, indicating a different conceptual framing of justice.

However, word-level embeddings have notable limitations. Many values are polysemous ("Macht" (power) as political power vs. ability to act), underrepresented in the corpus, or rarely used in their abstract psychological sense. Thus, while word-based embeddings offer a first approximation of value meaning, they must be supplemented by methods that move beyond single-word usage.

Sentence Level based Embedding

Therefore, to obtain richer representations of values, we construct sentence-level embeddings based on the formal definitions and conceptual descriptions of the values provided in Schwartz's own work. These definitions articulate the motivational goals, normative orientations, and evaluative priorities that constitute each value, and therefore serve more as a theory-aligned semantic anchor rather than an empirical usage-based embedding.

Each value definition consists of a short textual description (e.g., justice as fairness, equality, and protection of rights). The complete value defintions can be seen in the appendix. Our goal is to capture this description into a single vector, while preserving its conceptual

meaning. To this end, we encode each definition using the Smooth Inverse Frequency (SIF) method proposed by Arora et al. (2017).

For SIF let a sentence (or here the definition for a value) s consist of words $w_1, \ldots, w_n$, and let $\mathbf{v}_{w_i} \in \mathbb{R}^d$ denote the pre-trained word embedding of word w_i. A naive sentence embedding would be computed as a simple average:

$$\mathbf{v}_s^{\text{avg}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v}_{w_i}.$$

However, (Arora et al. 2017) show that such unweighted averaging is dominated by high-frequency function words (e.g., *the*, *and*, *of*), which contribute little semantic information but exert a disproportionate influence on the resulting vector.

To address this issue, SIF introduces a smooth inverse frequency weighting scheme. Each word is weighted by a factor inversely proportional to its corpus frequency:

$$\alpha(w) = \frac{a}{a + p(w)},$$

where $p(w)$ is the unigram probability of word w in a large reference corpus, and a is a small smoothing constant (typically $a \in [10^{-4}, 10^{-3}]$).

The weighted sentence embedding is then computed as:

$$\tilde{\mathbf{v}}_s = \frac{1}{n} \sum_{w \in s} \frac{a}{a + p(w)} \mathbf{v}_w.$$

This formulation has two important properties. First, it strongly downweights very frequent words without discarding them entirely, in contrast to hard stopword removal. Second, it preserves the contribution of moderately frequent but semantically meaningful terms-precisely those that characterize abstract values such as justice, equality, or responsibility. The result is a dense vector representation that captures the semantic content of the value definition while reducing noise from high-frequency words.

Although SIF weighting substantially improves sentence representations, Arora et al. (2017) show that sentence embeddings still tend to share a dominant common component that reflects generic properties of language, such as syntactic structure or corpus-specific regularities, rather than semantic content.

To remove this shared component, we perform principal component analysis (PCA) over the set of sentence embeddings and identify the first principal component $\mathbf{u}_1$. Each sentence embedding is then orthogonalized with respect to this direction:

$$\mathbf{v}_s = \tilde{\mathbf{v}}_s - \mathbf{u}_1 \quad \mathbf{u}_1^\top \tilde{\mathbf{v}}_s \quad .$$

This step removes the projection of the sentence vector onto the most dominant common direction, yielding a denoised embedding that better reflects sentence-specific semantic information.

The resulting vector $\mathbf{v}_s$ constitutes the final sentence-level embedding of a value definition. Ideally, this vector captures the conceptual meaning of the value as defined in psychological theory, rather than the empirical distribution of the value label in political discourse.

These value embeddings will then be compared for example using cosine similarity to embeddings of our political terms. Sentence-level embeddings thus enable the detection of value expressions even when value terms are not explicitly mentioned, grounding empirical text analysis in an externally defined semantic representation.

An very important methodological note is that the initial word embeddings used to construct the sentence-level embeddings are not drawn from any of the three ideology-specific models. Instead, we rely on a pre-trained, ideology-neutral baseline embedding model trained on the German Wikipedia (Yamada et al. 2020)The baseline model corresponds to a 300-dimensional *Wikipedia2Vec* embedding trained on the German Wikipedia dump from 2018 and obtained via the Hugging Face model repository.

When we measure the alignment between the embedding of a political term such as "Gerechtigkeit" (justice) and the sentence-level embedding of a value such as "Macht" (power) using cosine similarity, both vectors must be expressed in the same vector space. However, the sentence-level value embeddings are constructed from an ideology-neutral baseline model, whereas the political term embeddings come from our ideology-specific models, which are not directly comparable because their coordinate systems are arbitrarily rotated. We therefore first align each ideology-specific embedding space to the baseline space using orthogonal Procrustes alignment a method which will be explained later. Only after this alignment step, cosine similarities between political terms and sentence-level value embeddings become meaningful.

Axis based Embedding

A third and conceptually central way of operationalizing values is to represent them as semantic dimensions rather than as isolated points in the embedding space. A semantic dimension is a direction in the vector space along which concepts can be ordered continuously. Instead of asking whether a political concept is similar to a value label, this approach asks to what extent a concept aligns with one pole of a value dimension relative to its opposite.

To illustrate this intuitively, consider the value "Macht" (power). Rather than representing "Macht" as a single vector corresponding to the word "Macht," an axis-based approach constructs a direction running from powerlessness to power. Political concepts such as "Staat" (state), "Ordnung" (order), or "Autorität" (authority) can then be projected onto this

axis and compared according to how strongly they align with one end or the other. This representation captures meaning as a gradient rather than a category, which is particularly appropriate for abstract psychological constructs.

Representing values as semantic dimensions better mirrors Schwartz's theoretical conception of values as directional motivational goals. In Schwartz's framework, values do not function as discrete labels but as orientations that guide evaluation, interpretation, and judgment. Objects, actions, or ideas can express a value to varying degrees, and different values can compete or trade off against one another.

Semantic axes reflect this structure by allowing political concepts to be placed at intermediate positions between opposing value orientations. For example, a concept such as "Gerechtigkeit" (justice) may align partially with "Universalismus" (universalism) and partially with "Sicherheit" (security), depending on how it is framed in discourse. Axis-based embeddings thus provide a natural way to operationalize values as continuous conceptual dimensions rather than fixed word meanings.

Let us shortly introduce the mathematical foundation. Formally, let $\mathbf{v}(w) \in \mathbb{R}^d$ denote the embedding vector of a word w. A semantic axis is constructed from a set of anchor pairs (a_i, b_i), where a_i represents one pole of a value dimension and b_i represents the opposing pole.

Each anchor pair defines an offset vector

$$\mathbf{o}_i = \mathbf{v}(a_i) - \mathbf{v}(b_i).$$

For example to capture "power" by this way we could use core anchor pair "powerful" and "powerless". The corresponding offset vector would then be:

$$\mathbf{o}_{\text{power}} = \mathbf{v}(\text{powerful}) - \mathbf{v}(\text{powerless})$$

We then try to aggregate these offset vectors into a single direction vector $\mathbf{D}$ that approximates the latent value dimension. Here we compared two aggregation methods:

Mean Offset Method:

$$\mathbf{D}_{\text{mean}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i.$$

PCA-Based Method: The offset vectors are stacked into a matrix

$$\mathbf{M} = [\mathbf{d}_1, \ldots, \mathbf{d}_n],$$

and Principal Component Analysis is applied. The first principal component is taken as the axis direction:

$$\mathbf{D}_{\text{PCA}} = \text{PC}_1(\mathbf{M}).$$

As the Mean Offset Method showed initially the most promise, we used that for the rest of the study. A political concept w can then be projected onto the axis using cosine similarity:

$$\text{alignment}(w, \mathbf{D}) = \cos\ \mathbf{v}(w), \mathbf{D}\ .$$

This yields a scalar value indicating how strongly the concept aligns with the value dimension.

While semantic axes are theoretically appealing, recent methodological research has shown that their construction is highly sensitive to anchor choice. Boutyline and Johnston (2023) demonstrate that many intuitively plausible axes are unreliable because their anchor pairs do not encode a consistent semantic relation.

A key contribution of Bouytlines and Johnstons work is the distinction between *synonymy, antonymy,* and *parallelism*. *Synonymy* refers to semantic similarity between words (e.g., "mächtig" and "stark" / powerfull and strong). *Antonymy* refers to oppositional meaning (e.g., "mächtig" vs. "machtlos" / powerfull and powerless). *Parallelism*, by contrast, is a geometric property: whether the offset vectors between anchor pairs point in approximately the same direction.

Neither *synonymy* nor *antonymy* alone guarantees a high-quality axis. Two antonym pairs may be conceptually valid yet encode different relations (e.g., physical strength versus institutional authority), leading to misaligned offsets. What matters instead is *parallelism*: all anchor offsets should represent the same underlying semantic contrast. Parallelism thus functions as a reliability criterion analogous to internal consistency in psychometrics.

Building on this insight, Boutyline and Johnston (2023) recommend constructing axes from diverse but conceptually aligned anchor sets rather than relying on a single antonym pair. Effective axes are build with a core opposition pair that directly express the target dimension, proximate synonyms that reinforce the same relation, indirect or extended expressions that instantiate the dimension in related domains, and concrete examples that ground the abstraction in recognizable objects or entities. This strategy increases *parallelism* by averaging over multiple instantiations of the same semantic contrast.

We follow these recommendations closely. For each value dimension, anchors are chosen to fit into these categories that reflect increasing semantic distance from the core opposition while maintaining conceptual coherence. For example, the value "Macht" (power) is operationalized using the core pair (e.g., "mächtig"–"machtlos" (powerful–powerless)), proximate synonyms (e.g., "stark"–"schwach" (strong–weak), "einflussreich"–"wirkungslos" (influential–ineffective), indirect extensions (e.g., "Prestige"–"Verachtung" (prestige– contempt), morphological variants and concrete examples (e.g., "Herrscher"–"Diener" (ruler–servant), "König"–"Bauer" (king–peasant)).

By combining these anchor types, the resulting axis captures a generalized concept of power rather than a narrow lexical contrast. Aggregation is performed using both the mean offset and PCA-based methods as mentioned; following Boutyline and Johnston

(2023), this dual approach helps isolate the dominant shared direction among anchors and mitigates noise introduced by less parallel pairs.

The anchor word embeddings used to construct the semantic axes are again taken from the baseline model not from any of the ideology-specific models. This ensures that each axis direction **D** is defined in a fixed vector space. Political term embeddings are therefore again first aligned into this baseline space via orthogonal Procrustes alignment before being projected onto the axes.

5.3 From Operationalization to Measurement of Semantic Change

Having operationalized both our political key terms and the value representations, we can now turn to the question of how semantic differences between words are actually measured. The analysis proceeds in two stages. In the first step, we set aside the value-based interpretation and look only at how the key political terms differ across ideological contexts in the embedding space itself, without relating them to any value dimensions. Only in the second step, we reintroduce the operationalized values - first as simple word embeddings, then as sentence based definitions, and finally as semantic axes - to assess how ideological variation in political language aligns with underlying value dimensions.

Semantic Change and Word Stability

As mentioned a core challenge in computational semantics is that semantic meaning itself is not directly observable and therefore cannot be measured independently of language use. Even the most sophisticated embedding models provide access only to relational representations of words, not to meaning as such. Consequently, computational approaches cannot directly detect semantic change in the strong linguistic sense, but only assess whether a word's usage remains stable across different contexts or corpora. Recent work therefore reframes the problem of semantic change in terms of word stability which describes the degree to which a word occupies a similar position in aligned embedding spaces trained on different corporas. Azarbonyad et al. (2017) for example explicitly adopt this perspective, defining semantic stability as similarity between a word's vector representations and treating instability as indirect evidence of semantic change rather than its direct measurement. Therefore when we speak of greater observed semantic change in one word compared to another, we mean that the former exhibits lower word stability than the latter.

Cosine Similarity

The most commonly used measure for assessing the stability of word through many corpora is cosine similarity. Cosine similarity quantifies the angle between two vectors in a high-dimensional space and is defined as:

$$\cos(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \, \|\mathbf{v}_2\|}.$$

Unlike distance measures that depend on vector length, cosine similarity captures directional similarity, which is where most semantic information is encoded in distributional word embeddings. Two vectors pointing in the same direction have a cosine similarity close to 1, while lower values indicate increasing divergence in contextual usage.

In studies of semantic change, cosine similarity is typically used to measure word stability by comparing a word's vector representations across aligned embedding spaces trained on different corpora. However word vectors trained from two different corporas can't be compared by cosine similarity without certain refinements. Word embeddings trained on different corpora define separate embedding spaces which means that each training run produces its own embedding space, and because these spaces are learned independently, they do not share a common coordinate system. Random initialization in models such as Word2Vec means that embeddings start from different weights, and during training the resulting dimensions acquire different semantic interpretations. Consequently, even embeddings trained with identical parameters may differ by an arbitrary rotation or reflection. While such transformations preserve relative similarities within each space, they prevent meaningful direct comparison of word vectors across spaces.

To make vectors from different models comparable, we apply Orthogonal Procrustes Alignment, a standard technique in diachronic embedding research (Hamilton et al. 2018). The idea is to find a rotation matrix R that maps one embedding space onto another using a shared set of anchor words. Given two matrices X and Y containing the vectors of the same anchor words from two embedding models, Procrustes alignment solves the following optimization problem:

$$\min_R \|XR - Y\| \quad \text{subject to} \quad R^\top R = I.$$

The orthogonality constraint ensures that distances and angles within the embedding space are preserved. Intuitively, the method rotates one space so that corresponding words in the two models align as closely as possible, without distorting the internal semantic structure. Then we can multiply a word vector by the rotation matrix and map it so into the coordinate system of the other embedding space.

Following common practice, we use a set of high-frequency words (the 500 most common words) shared across both corpora as anchors, as these words tend to have relatively stable meanings and well-estimated vectors. Once the alignment is applied, vectors for the same word can be directly compared across embedding spaces, making cosine similarity a meaningful measure of word stability across ideological viewpoints.

Once the embedding spaces can be be aligned with the rotation matrix, word stability can be measured by directly comparing the same word across ideological corpora. Concretely, we train one Word2Vec model on right-wing parliamentary speeches and another on left-wing speeches, align with orthogonal Procrustes alignment, and then extract the vector for a given word -such as "Gerechtigkeit" (justice)- from each model. Word stability is then computed as the cosine similarity between the aligned vectors:

$$\text{stability}(\textit{Gerechtigkeit}) = \cos\ \mathbf{v}^{\text{rightwing}}_{\textit{Gerechtigkeit}}, \mathbf{v}^{\text{leftwing}}_{\textit{Gerechtigkeit}} \quad .$$

A high cosine similarity indicates that "Gerechtigkeit" (justice) occupies a similar position in both embedding spaces and is therefore used in comparable contextual environments across the two ideological corpora. A lower similarity, by contrast, suggests systematic differences in usage, which we interpret as reduced word stability and increased semantic divergence.

Neighbor based Approaches

While cosine similarity compares the position of a word directly across embedding spaces, an alternative and complementary way to assess semantic difference is to examine changes in a word's local neighborhood. Neighbor-based approaches focus on how the set of words most closely associated with a target term differs across ideological contexts. The underlying intuition is that meaning is reflected not only in a word's absolute position in the embedding space, but also in the company it keeps.

Once the embedding spaces have been created after the training we identify for each target word its k nearest neighbors in each ideological embedding space, based on cosine similarity. For a given political term - such as "Gerechtigkeit" (justice) - this yields two neighbor sets: one derived from right-wing discourse and one from left-wing discourse. Semantic difference is then quantified by comparing the overlap between these two sets.

Formally, let $N_k^{(Right)}(w)$ and $N_k^{(Left)}(w)$ denote the sets of the k nearest neighbors of word w in the conservative and progressive embedding spaces, respectively. A simple and widely used measure of neighborhood stability is the Jaccard similarity:

$$\text{neighbor\_stability}(w) = \frac{N_k^{(Right)}(w) \cap N_k^{(Left)}(w)}{N_k^{(Right)}(w) \cup N_k^{(Left)}(w)}.$$

High overlap indicates that a word is surrounded by largely the same semantic associates across ideological contexts, suggesting stable usage. Low overlap, by contrast, indicates that different concepts are activated around the same word in different ideological corpora, which we interpret as semantic divergence.

Neighbor-based measures have several important advantages. First, they are locally interpretable: changes in a word's neighbors can be inspected directly, allowing researchers to see how meanings differ rather than merely that they differ. For example, if "Gerechtigkeit" (justice) is surrounded by terms such as "Ordnung" (order), "Sicherheit" (security), and "Strafe" (punishment) in conservative discourse, but by "Gleichheit" (justice), "Solidarität" (solidarity), and "Teilhabe" (inclusion) in progressive discourse, this provides a concrete illustration of ideological framing differences.

Second, neighbor-based approaches are less sensitive to global shifts in embedding space geometry than direct vector comparisons. Even if a word's absolute position remains relatively stable, changes in its nearest neighbors can reveal subtle but meaningful shifts in interpretation. This makes neighborhood analysis particularly useful for detecting nuanced semantic differences that may not result in large cosine-distance changes.

Third, this method aligns closely with the distributional hypothesis underlying word embeddings. If meaning is derived from contextual co-occurrence, then examining a word's immediate semantic neighborhood provides a direct window into its contextual usage. As shown by prior work on semantic shift and viewpoint-specific embeddings, neighborhood change is a reliable indicator of meaning variation across corpora (Azarbonyad et al. 2017; Hamilton et al. 2018).

In this thesis, neighbor-based measures are used as a complement to cosine-based word stability. While cosine similarity captures *global positional differences*, neighbor-based measures overlap captures *local contextual restructuring*. Taken together, these measures provide a more complete picture of how political concepts vary semantically across ideological viewpoints.

Combination of both Methods

While cosine similarity and neighbor overlap capture complementary aspects of semantic stability, each measure on its own provides only a partial view. Cosine similarity reflects global positional stability that is, whether a word occupies a similar region of the embedding space across ideological corpora, whereas neighbor-based measures capture local contextual stability, namely whether a word is embedded among similar semantic

associates. To obtain a single, interpretable measure of semantic stability, we therefore combine both signals into a unified stability score.

The procedure proceeds as follows. For each target word, we first compute a neighbor-based difference score, which quantifies how strongly the word's local neighborhood differs across the two embedding spaces. This score is then transformed into a stability measure by subtracting it from one, so that higher values consistently indicate greater stability. In parallel, we compute the cosine similarity between the aligned word vectors and rescale it from the interval $[-1, 1]$ to $[0, 1]$, ensuring comparability between the two components.

These two normalized quantities, global positional stability and local neighborhood stability, are then combined using a weighted aggregation scheme. We considered three aggregation methods: the arithmetic mean, the geometric mean, and the harmonic mean. Each method reflects a slightly different theoretical emphasis. The arithmetic mean treats both components as additive and allows high values in one dimension to compensate for lower values in the other. The geometric mean, used as the default in this study, balances the two components multiplicatively and reduces the influence of extreme values, ensuring that both positional and neighborhood stability contribute meaningfully to the final score. The harmonic mean, finally, penalizes low values more strongly and thus highlights cases in which instability in either component signals substantial semantic divergence.

A weighting parameter $\alpha \in [0, 1]$ controls the relative contribution of the two components. In this thesis, $\alpha = 0.5$ assigns equal importance to positional and neighborhood-based stability, reflecting the assumption that semantic meaning is jointly determined by global placement and local contextual relations. The resulting stability score ranges from 0 to 1, with higher values indicating more stable usage of a word across ideological corpora.

By combining cosine similarity and neighbor-based measures into a single score, this approach provides a more robust and nuanced estimate of semantic stability than either metric alone. It allows us to rank political key terms according to their degree of semantic divergence while retaining sensitivity to both broad shifts in meaning and finer-grained contextual reconfigurations.

5.4 Training Data and Corpus Construction

Now that we have outlined how we operatonalize the political terms and Schwartz's value theory and how we measure semantic change (word stability) we now turn to the empirical foundation of the analysis: the data on which the embedding models are trained.

The training data used in this thesis consists of parliamentary speeches from two institutional settings: the German Bundestag and the European Parliament. Specifically, the

corpus includes all plenary speeches held in the Bundestag during the last three legislative periods (19th, 20th, 21th), as well as speeches delivered by German members of the European Parliament during the 9th and 10th legislative terms. These time frames were chosen deliberately, as they cover the period in which the *Alternative für Deutschland* (AfD) is present as a parliamentary party and thus constitutes the only clearly german right-wing ideological bloc within both institutions. This temporal restriction is unfortunately limited but necessary for constructing a meaningful right-wing corpus that is comparable to left-wing and centrist discourse.

All speeches are annotated with metadata identifying the party affiliation of the speaker. Based on this information, speeches were grouped into three ideological corpora: left-wing (S*PD, Bündnis 90/Die Grünen, Die Linke*), centrist (*CDU/CSU*), and right-wing (*AfD*). While the precise ideological placement of the parties can be debated, a more fine-grained classification falls outside the scope of this thesis; we therefore adopt this broad and commonly used categorization for the sake of analytical clarity. For each ideological group, speeches from the Bundestag and the European Parliament were combined to form a single training corpus.

In total, the resulting corpora comprise several million tokens per ideological group. The left-wing corpus contains approximately 10.8 million tokens, the centrist corpus approximately 8.0 million tokens, and the right-wing corpus approximately 3.5 million tokens.

The corpora also differ in vocabulary size. Using lemmatized forms, the left-wing corpus contains approximately 55.900 unique tokens, the centrist corpus around 46.900, and the right-wing corpus about 34.700. These differences may partly reflect linguistic diversity across ideological camps; however, the primary reason for the variation in vocabulary size is that the right-wing corpus consists exclusively of speeches from a single party (AfD). Such asymmetries in vocabulary size are a known challenge in comparative embedding studies, as smaller corpora may yield noisier representations for low-frequency words. However, this issue is mitigated in several ways: by restricting the analysis to sufficiently frequent political key terms, by using aligned embedding spaces, and by relying on relative rather than absolute semantic comparisons. As a result, differences in vocabulary size are acknowledged but remain manageable for the purposes of this study.

Parliamentary speeches are particularly well suited as training data for this research. First, they represent elite political discourse, where ideological positions are articulated explicitly and strategically. Speakers are incentivized to frame political concepts in ways that align with their party's values and worldview, making this domain especially appropriate for studying ideological semantic variation. Second, the language used in parliamentary debate is both formalized and recurrent, which supports the reliable estimation of distributional semantics. Third, the availability of high-quality metadata allows for a clean and transparent ideological partition of the corpus, avoiding the need for automated ideology classification or proxy labels.

Several extensions of this data choice are possible. Future work could include party manifestos, media appearances, or voter-generated texts such as social media posts to capture a broader range of political communication. However, such data would introduce additional noise and complicate ideological classification. Given the aims of this thesis the use of parliamentary speeches from the Bundestag and European Parliament provides a well-balanced trade-off between interpretability, corpus size, and methodological control.

5.5 Summary of the Analytical Pipeline

To conclude the methodological section, we briefly summarize the analytical pipeline used in this thesis. The analysis proceeds in a series of clearly defined steps that connect theoretical assumptions about meaning and values to empirical measurements in embedding space.

First, parliamentary speech data from the German Bundestag and the European Parliament are collected and partitioned into three ideological corpora, left-wing, centrist, and right-wing, based on the party affiliation of the speaker. Separate Word2Vec embedding models are then trained on each ideological corpus, resulting in distinct semantic spaces that capture how political language is used within each ideological context.

Second, the semantic stability of political key terms is measured using two complementary approaches: cosine similarity between orthogonal procruste aligned word vectors, which captures global positional differences, and neighbor-based overlap measures, which capture changes in local semantic neighborhoods. These measures are combined into a single stability score that reflects both global and local aspects of semantic change.

Third, we operationalize Schwartz's values in three ways: (i) value labels (word-based), (ii) sentence-level embeddings derived from Schwartz's value definitions, and (iii) semantic axes constructed from anchor pairs. For the sentence- and axis-based approaches, we rely on an ideology-neutral baseline embedding model (*Wikipedia2Vec*) to construct the value embeddings. Political key terms are then mapped into this baseline space and projected onto the resulting value embeddings to examine how ideological differences in language align with underlying value dimensions.

Finally, the resulting stability scores and value alignments are analyzed comparatively across ideological groups. This enables an empirical assessment of whether and how political ideology shapes the semantic meaning of core political concepts, and whether these differences correspond to value-based patterns predicted by political psychology.

For completeness and reproducibility, we report the main training parameters used for our three Word2Vec models. Unless stated otherwise, the same configuration was applied to all ideological corpora to ensure comparability across embedding spaces.

Table 5.1: Word2Vec training parameters used across all models.

Parameter	**Value**
Vector size	300
Context window	5
Minimum word frequency	3
Architecture	Skip-gram (sg = 1)
Negative samples	15
Subsampling rate	0.0001
Training epochs	10

6 Results

Now we can finally present the empirical results of the analytical pipeline described above. We begin by examining the semantic stability of predefined key terms across ideological corpora. We then turn to local, neighbor-based analyses that illustrate how ideological differences manifest in concrete linguistic contexts. Finally, we relate observed semantic differences to Schwartz's value dimensions to evaluate whether ideological variation in political language aligns with theoretically expected value structures.

6.1 Overall Semantic Stability of Political Terms

6.1.1 Stability Scores of Key Terms

Table 6.1 reports the stability scores for the predefined political terms across ideological corpora. Overall, the results indicate substantial variation in stability both across and within the three conceptual categories: Institutional, Policy, and Cultural.

Table 6.1: Stability scores by term and category

category	term	stability
Institutional	grundgesetz	0.385
Institutional	demokratie	0.366
Institutional	verfassung	0.275
Institutional	freiheit	0.271
Institutional	rechtsstaat	0.253
Institutional	bundestag	0.234
Institutional	opposition	0.149
Institutional	sozialstaat	0.133
Institutional	bundesregierung	0.129
Institutional	föderalismus	0.121
Policy	inflation	0.352
Policy	migration	0.286
Policy	asyl	0.264
Policy	frieden	0.244
Policy	umweltschutz	0.220
Policy	sicherheit	0.218
Policy	klimaschutz	0.214
Policy	integration	0.212
Policy	energiewende	0.179
Policy	gerechtigkeit	0.172
Cultural	europa	0.253
Cultural	würde	0.206
Cultural	identität	0.205
Cultural	volk	0.202

Continued on next page

Table 6.1: Stability scores by term and category

category	term	stability
Cultural	heimat	0.190
Cultural	solidarität	0.180
Cultural	deutschland	0.155
Cultural	patriotismus	0.130
Cultural	nation	0.111
Cultural	leitkultur	0.002

As expected institutional terms display comparatively high stability on average, with "Grundgesetz" (basic law) (0.385) and "Demokratie" (democracy) (0.366) exhibiting the highest stability scores within the entire set. This suggests that core constitutional concepts tend to occupy relatively similar semantic positions across the different ideological contexts. At the same time, the lower-ranked institutional terms - such as "Föderalismus" (federalism) (0.121), "Bundesregierung" (government) (0.129), and "Sozialstaat" (welfare state) (0.133) - show noticeably reduced stability. This broad spread of stability values within this category suggests a split between more stable constitutional references like "Grundgesetz" (basic law) (0.385) and more politicized terms that invite contestation, with actors attempting to reframe them. For instance, the opposition can benefit from negative portrayals of the "Bundesregierung" (government), where as the government parties try to frame it as something positive. Policy-related terms occupy an intermediate position in terms of overall stability. While concepts such as "Inflation" (inflation) (0.352) and "Migration" (migration) (0.286) exhibit a surprising relative high stability, others including "Gerechtigkeit" (justice) (0.172), "Energiewende" (energy transition) (0.179), and "Integration" (integration) (0.212) show markedly lower scores.

The Cultural category shows the lowest aggregate stability, a result that is largely driven by extreme outliers. In particular, "Leitkultur" (guiding culture) (0.002) exhibits an almost complete lack of stability, indicating near-total divergence in its semantic meaning across ideological embeddings. This finding aligns with the terms expected roles as highly contested symbolic concepts. Other cultural terms such as "Nation" (nation) (0.111), "Patriotismus" (patriotism) (0.130), and "Deutschland" (Germany) (0.155) also display comparatively low stability, whereas more abstract or universally framed concepts like "Europa" (Europe) (0.253) occupy more central and stable positions. The pronounced instability of culturally charged terms suggests that symbolic and identity-related language is especially sensitive to ideological context.

Table 6.2: Sum of stability scores by category

category	stability
Institutional	2.317
Policy	2.361
Cultural	1.635

Table 6.2 summarizes these patterns and confirms that Institutional and Policy terms exhibit comparable aggregate stability (2.317 and 2.361), while Cultural terms are substantially less stable overall (1.635). Taken together, these results indicate that ideological semantic variation is not evenly distributed but concentrates on culturally and normatively loaded concepts, whereas constitutional and policy related language remains relatively stable across ideological divides. This supports the interpretation that ideological differences in political language are more strongly expressed through symbolic and value laden vocabulary than through references to formal institutions or broadly shared policy concerns.

6.1.2 Local Semantic Change: Neighbor-Based Analysis of Key Terms

Table 6.3 lists the Neighbor-based difference scores for our political terms, showing how their local semantic neighborhoods diverge across ideological embeddings. Unlike earlier aggregated stability scores, this zooms in on local context for a finer view. For each key term w, we retrieve its top-100 nearest neighbors in each embedding and, for every model pair (i, j) compute a neighbor–difference score using the Jaccard distance. With three embeddings, we then calculate the mean jacard difference: $\bar{d}(w) = \frac{d_{1,2}(w)+d_{1,3}(w)+d_{2,3}(w)}{3}$.

Table 6.3: Neighbor difference by term and category

category	term	neighbor difference	normalized neighbor difference
Institutional	bundesregierung	0.981	0.958
Institutional	föderalismus	0.980	0.948
Institutional	sozialstaat	0.974	0.915
Institutional	opposition	0.967	0.871
Institutional	bundestag	0.940	0.703
Institutional	freiheit	0.921	0.586
Institutional	rechtsstaat	0.913	0.535
Institutional	verfassung	0.898	0.447
Institutional	demokratie	0.854	0.176
Institutional	grundgesetz	0.830	0.028
Policy	gerechtigkeit	0.958	0.816

Continued on next page

Table 6.3: Neighbor difference by term and category

category	term	neighbor difference	normalized neighbor difference
Policy	sicherheit	0.947	0.749
Policy	energiewende	0.943	0.720
Policy	integration	0.934	0.669
Policy	klimaschutz	0.931	0.649
Policy	umweltschutz	0.930	0.643
Policy	frieden	0.921	0.585
Policy	asyl	0.896	0.433
Policy	migration	0.873	0.292
Policy	inflation	0.826	0.000
Cultural	leitkultur	0.988	1.000
Cultural	nation	0.981	0.958
Cultural	patriotismus	0.974	0.915
Cultural	deutschland	0.973	0.904
Cultural	würde	0.953	0.783
Cultural	solidarität	0.949	0.758
Cultural	heimat	0.947	0.749
Cultural	volk	0.941	0.713
Cultural	identität	0.940	0.703
Cultural	europa	0.926	0.619

Table 6.4: Sum of neighbor differences by category

Category	Sum of neighbor differences
Institutional	9.258
Policy	9.159
Cultural	9.572

Table 6.4 largely mirrors the stability patterns observed in Table 6.2. Terms that exhibited low overall stability in the stability score, most notably culturally and normatively charged concepts, also show the strongest divergence in their immediate semantic neighborhoods. Conversely, terms that ranked as comparatively stable in the earlier analysis tend to retain more overlap in their closest associates. However, it should be noted that neighborhood difference scores are very high across the entire set of political terms, which limits the evaluative strength of the analysis and means that the conclusions hold only in relative terms.

6.1.3 Semantic Neighborhood Visualizations

The semantic neighborhood visualizations provide a complementary, more intuitive perspective on the patterns observed in the previous tables. While the above analyses capture the degree of divergence in numerical values, the t-SNE projections makes this divergence visible with concrete words. It is important to note that the visualized neighborhoods do not necessarily correspond to the nearest neighbors. This discrepancy is an expected consequence of the t-SNE algorithm, which prioritizes the preservation of local neighborhood relations while allowing global distances and exact rank orderings to be distorted. As a result, words that are among the closest neighbors in the high-dimensional embedding space may appear slightly farther apart in the two-dimensional projection, and vice versa. The visualization should therefore be interpreted qualitatively rather than as a precise geometric mapping. For reasons of readability and interpretability, we present only four selected examples that offer the clearest qualitative insights; displaying all semantic neighborhoods would be prohibitively lengthy and many cases are not readily interpretable. The exact method for creating our neighbor space is explained in the appendix.

Asyl

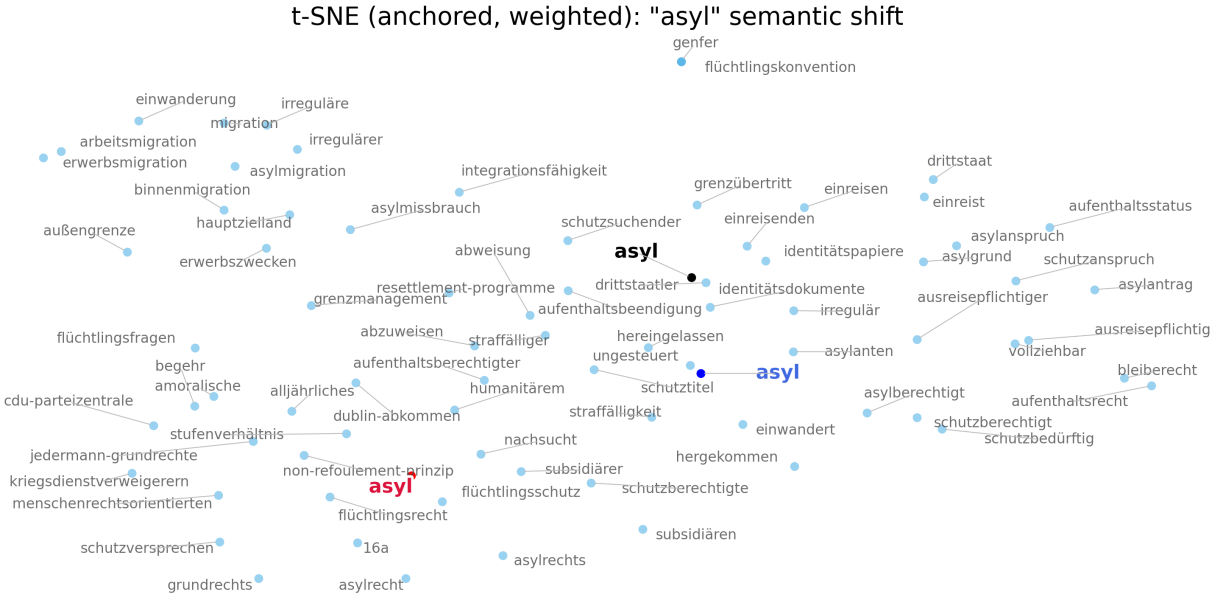

Figure 6.1: Anchored and weighted t-SNE visualization illustrating how the semantic neighborhood of the term "asyl" (asylum) shifts across ideological contexts. Each point represents a neighboring term in the embedding space, with spatial proximity indicating higher semantic similarity. The highlighted anchor points correspond to the term "asyl" as embedded in different ideological models: left-wing, right-wing, and center (black). Surrounding clusters reflect dominant contextual associations, ranging from humanitarian and legal framings to administrative and control-related perspectives.

The first example in which the ideological separation of semantic neighborhoods becomes particularly salient is the case of the term "Asyl" (asylum). In the projected space (figure 6.1), left-wing usages cluster tightly around terms associated with humanitarian protection and moral obligation, such as "Flüchtlingsrecht" (refugee law), "Flüchtlingsschutz" (refugee protection), "Asylrecht" (asylum law), "Menschenrechtsorientierten" (human-rights-oriented), or "humanitärem" (humanitarian). These neighbors indicate that "Asyl" (asylum) is predominantly embedded in a frame that emphasizes vulnerability, rights, and international responsibility. The relative compactness of this cluster further suggests a fairly coherent and consistent framing across left-wing discourse, in which asylum is treated primarily as a normative and humanitarian issue. By contrast, the right-wing and center neighborhood forms a clearly distinct cluster that is oriented toward legal, administrative, and security-related concepts. Here, "Asyl" (asylum) is situated near terms associated with "Identitätsdokumente" (identity documents), "Identitätspapiere" (identity papers), "irregulär" (irregular), and "Straffälligkeit" (criminality). Rather than foregrounding the individuals seeking asylum, this contextual environment frames the concept in terms of regulation, enforcement, and potential strain on state capacity. Note, too, that the center and right-wing neighborhoods are closer to each other than to the

left-wing one. To make the differences between the models more tangible, let us turn to concrete example sentences, listed in table 6.5, involving the term "Asyl" (asylum) on which the respective models were trained:

Model	Example sentence
Right-wing	*"Dann lassen Sie mich eines klarstellen: Der Begriff ist falsch; denn der Großteil der Menschen, die hierzulande **Asyl** beantragen, ist weder asylberechtigt noch vor etwas geflohen, sondern sie werden angelockt von unseren Sozialleistungen, von denen sie besser leben können als im Heimatland." (Martin Sichert, AfD, 19.01.2025)*
Center	*"Die Grundproblematik ist: Um die Akzeptanz für das Grundrecht auf **Asyl** zu erhalten, müssen wir alles daransetzen, dass diejenigen, die aus missbräuchlichen Gründen kommen, konsequenterweise wieder abgeschoben werden." (Hans-Jürgen Irmer, CDU, 23.10.2019)*
Left-wing	*"Liebe Griechen, das ist unsere gemeinsame Grenze, es ist ein Bruch des Europarechts, des Völkerrechts, der Genfer Flüchtlingskonvention, wenn man den Zugang zu **Asyl** aussetzt und schutzsuchende Menschen mit unverhältnismäßiger Gewalt beschießt und zurückdrängt." (Luise Amtsberg, Grüne, 19.01.2018)*

Table 6.5: Representative example sentences containing the term "Asyl" (asylum) drawn from the training corpora of the right-wing, center, and left-wing models. The examples illustrate how the semantic framing of asylum differs across ideological contexts.

Inflation

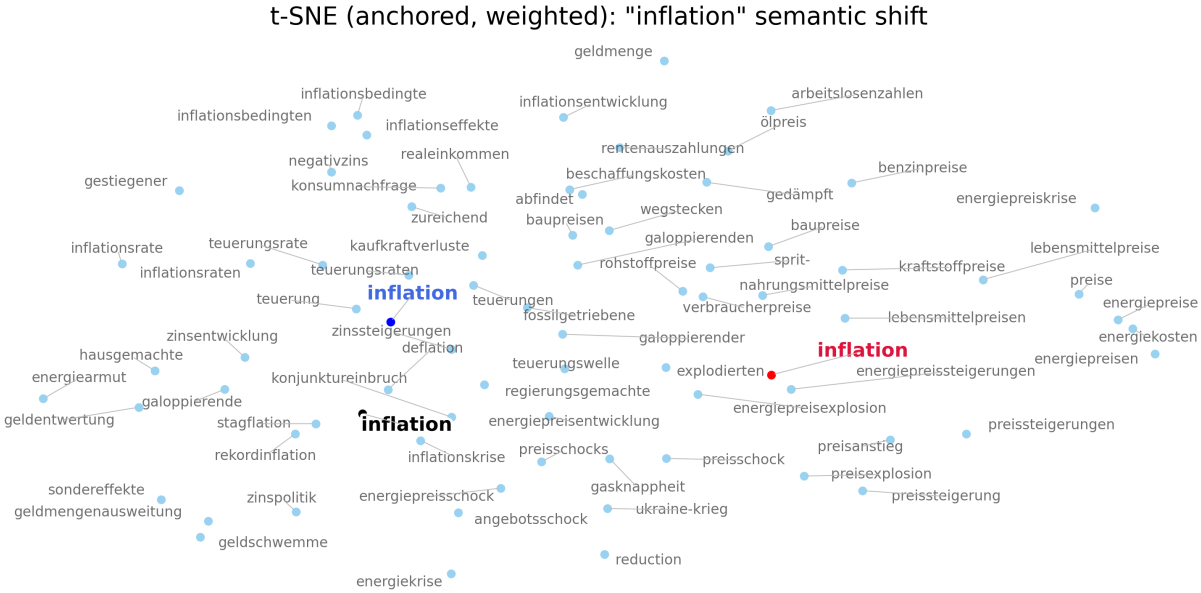

Figure 6.2: Anchored and weighted t-SNE visualization illustrating how the semantic neighborhood of the term "inflation" (inflation) shifts across ideological contexts. Each point represents a neighboring term in the embedding space, with spatial proximity indicating higher semantic similarity. The highlighted anchor points correspond to the term "inflation" as embedded in different ideological models: left-wing, right-wing, and center (black). Surrounding clusters reflect dominant contextual associations, including monetary policy, price dynamics, and energy-related shocks.

A different picture emerges for "Inflation" (inflation), which appears markedly less ideologically segmented in the semantic neighborhood visualization (figure 6.2). Unlike "Asyl" (asylum), the projected neighborhoods for "Inflation" (inflation) show semantic overlap across ideological contexts, with all three clusters gravitating toward economically technical and descriptive terms. Where differences do appear, they are comparatively subtle and concern emphasis rather than framing. Left-wing discourse tends to embed "Inflation" (inflation) somewhat closer to terms related to social consequences, such as "Lebensmittelpreise" (food prices), "Nahrungsmittelpreise" (food prices), or "Preisanstieg" (price increase), whereas right-wing and center usage shows slightly stronger proximity to market-oriented or fiscal terminology like "Zinssteigerungen" (interest rate increases), "Konjunktureinbruch" (economic downturn), or "Zinsentwicklung" (interest rate development). This visual pattern helps explain why Inflation ranks among the most stable policy terms in the earlier analyses (see 6.1). The semantic neighborhoods suggest that ideological actors largely agree on what Inflation refers to and how it is discussed, even if they disagree on its causes or remedies at the level of argumentation. Table 6.6 again lists concrete examples of the usage of "Inflation" (inflation) in the different corpora:

Model	Example sentence
Right-wing	*Es gibt nur zwei Lösungen: entweder noch höhere Steuern, wie es die SPD heute vorschlägt, oder eine Ausweitung der Geldmenge, bei der die Bürger durch Entwertung des Geldes in Form einer **Inflation** zahlen. So oder so sind immer die fleißigen Werktätigen die Dummen." (Martin Sichert, AfD, 15.05.2020)*
Center	*"Natürlich ist für die **Inflation** auch die Geldpolitik der Europäischen Zentralbank, die zu Recht unabhängig ist, entscheidend. Aber der Handlungsspielraum der Europäischen Zentralbank, die selbst in Zeiten dieser **Inflation** noch auf negative Einlagezinsen setzt und Anleihekäufe tätigt, ergibt sich natürlich auch aus der Fiskalpolitik der Staaten." (Thorsten Frei, CDU, 01.06.2022)*
Left-wing	*"Das sind nun einmal eine Energiepreiskrise und eine **Inflation**, die den Menschen den Boden unter den Füßen wegziehen. Die Menschen zahlen bei jedem Wocheneinkauf dick obendrauf." (Christian Leye, Linke, 20.05.2022)*

Table 6.6: Representative example sentences containing the term "Inflation" (inflation) drawn from the training corpora of the right-wing, center, and left-wing models. The examples illustrate how the semantic framing of Inflation differs across ideological contexts.

Patriotismus

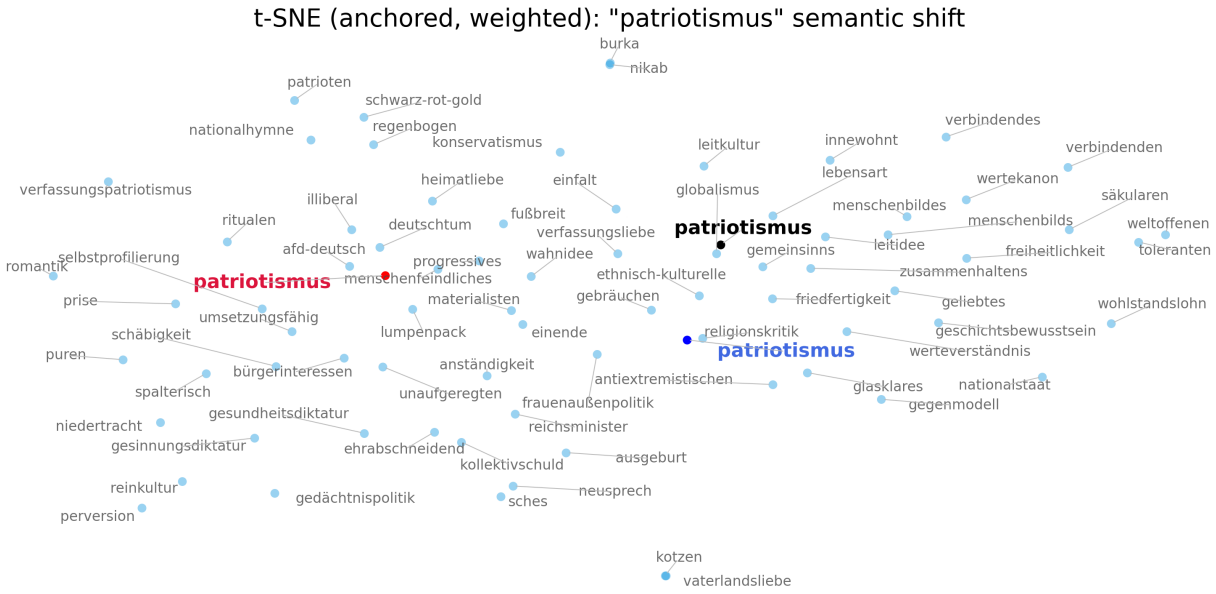

Figure 6.3: Anchored and weighted t-SNE visualization illustrating how the semantic neighborhood of the term "patriotismus" (patriotism) shifts across ideological contexts. Each point represents a neighboring term in the embedding space, with spatial proximity indicating higher semantic similarity. The highlighted anchor points correspond to the term "patriotismus" as embedded in different ideological models: left-wing, center (black), and right-wing. Surrounding clusters reflect contrasting interpretations, ranging from inclusive civic conceptions to exclusionary and identity-based associations.

In our next example "Patriotismus" (patriotism), the visualization reveals a coherent ideological structuring of the term (figure 6.3). In the projected space, the right-wing and the center usage of "Patriotismus" (patriotism) forms a cluster with references to national belonging and symbolic affirmation. Terms like "Gemeinsinns" (sense of community), "Friedfertigkeit" (peacefulness), "Lebensart" (way of life), or "Gebräuchen" (customs) emphasize collective identity, continuity, and positive valuation, indicating that patriotism is used as a normative concept with affirmative connotations.
In contrast, the left-wing semantic neighborhood is markedly more negative, with terms like "Menschenfeindliches" (anti-human), "Lumpenpack" (rabble), or "illiberal" (illiberal) that problematize the term or contextualize national attachment rather than celebrating it. In left-wing discourse, "Patriotismus" (patriotism) is frequently invoked in critical, reflective, or delimiting ways rather than as a positive identity marker.
Overall, the visualization makes it clear that "Patriotismus" (patriotism) carries very different meanings depending on ideological context. These differences are not subtle shifts in wording but reflect distinct ways of embedding the term into broader narratives, which helps explain why it shows such a low stabilty score (see 6.1). Table 6.7 again lists concrete examples of the usage of "Patriotismus" (patriotism) in the different corpora:

Model	Example sentence
Right-wing	*"Als AfD-Fraktion verfolgen wir das Prinzip eines solidarischen **Patriotismus** und keinen Globalismus." (Hannes Gnauck, AfD, 01.12.2022)*
Center	*"Deswegen fordern wir eine verstetigte Aufwertung des Tags des Grundgesetzes und auch einen ganzjährigen Einsatz für einen verbindenden **Patriotismus** in unserem Land." (Phillip Amthor, Union, 16.05.2024)*
Left-wing	*"Wer mehr **Patriotismus** und nationale Symbolik gegen Rechtsextremismus fordert, der will Feuer mit Benzin löschen." (Janine Wissler, Linke, 24.05.2023)*

Table 6.7: Representative example sentences containing the term "Patriotismus" (patriotism) drawn from the training corpora of the right-wing, center, and left-wing models. The examples illustrate how the semantic framing of patriotism differs across ideological contexts.

Bundestag

As for our last example, "Bundestag" provides a case of a term with very high semantic stability across the ideological embedding models. As the visualization shows (figure 6.4), the nearest neighbors of "Bundestag" cluster around procedural, institutional, and organizational references such as "Plenum" (plenary session), "Parlament" (parliament), "Haushaltsausschuss" (budget committee) and legislative bodies. These associations are largely shared across ideological contexts, indicating that the term is consistently embedded in a common institutional frame. Table 6.8 lists again concrete examples from the corpora

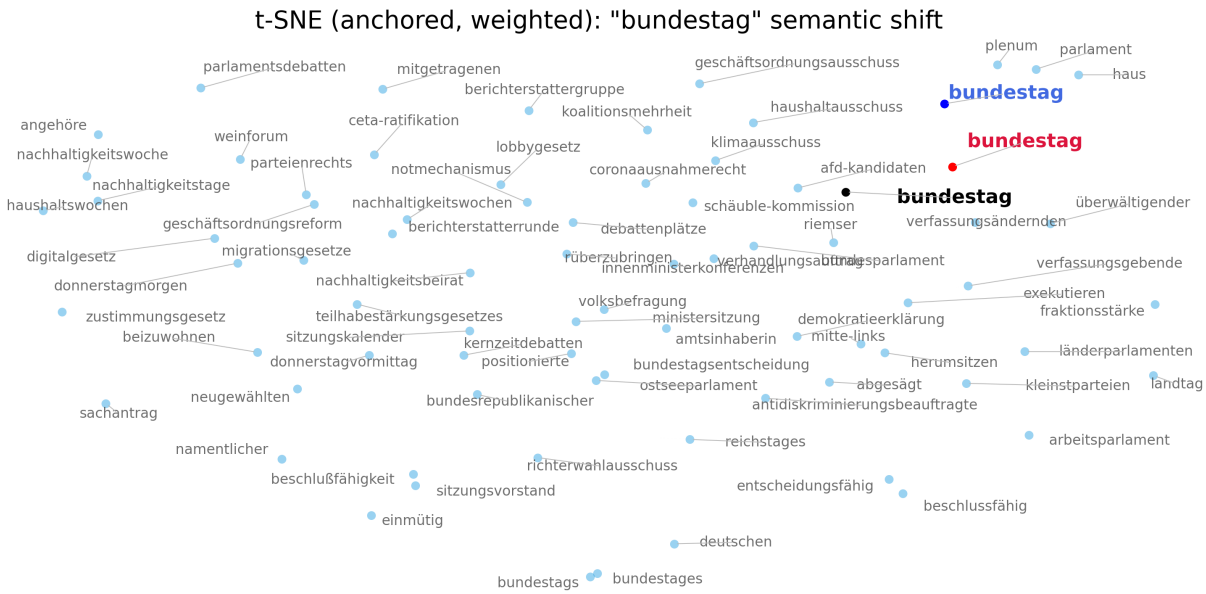

Figure 6.4: Anchored and weighted t-SNE visualization illustrating how the semantic neighborhood of the term "Bundestag" shifts across ideological contexts. The highlighted anchor points correspond to the term "Bundestag" as embedded in different ideological models: left-wing, center (black), and right-wing.

Model	Example sentence
Right-wing	*"Ich würde mir wünschen, dass sich der Deutsche **Bundestag** hier zum Bewahrer der deutschen Sprache erklärt, sich als Hüter der verbalen Kommunikation mit dem Volk versteht und wir uns nicht mit überbordendem Snobismus einer Sprache einer abgehobenen Nomenklatura befleißigen." (Rainer Kraft, AfD, 25.03.2021)*
Center	*"Wir als Deutscher **Bundestag** haben die Zügel ganz fest in der Hand. Bei dem, was wir heute machen und was wir im letzten November gemacht haben, als wir § 28a Infektionsschutzgesetz auf den Weg gebracht haben, geht es darum, dass die wesentlichen Entscheidungen, die uns das Bundesverfassungsgericht mit dem Kollegen Stephan Harbarth als Präsident aufgetragen hat, wir als Deutscher **Bundestag**, wir als Parlament selber treffen. Und wir machen das genau so, wie die Verfassung das vorsieht, liebe Kolleginnen und Kollegen. Wir treffen hier im Deutschen **Bundestag** die wesentlichen Entscheidungen." (Dr. Jan-Marco Luczak , Union, 12.02.2021)*
Left-wing	*"Deshalb wollen wir als Deutscher **Bundestag** der Bundeskanzlerin Rückenwind geben; so kann man es nennen. Wir müssen aber auch Druck ausüben, um interfraktionell dieses gemeinsame Ziel zu erreichen." (Steffi Lemke, Bündnis 90/Die Grünen, 08.10.2020)*

Table 6.8: Representative example sentences containing the term "Bundestag" drawn from the training corpora of the right-wing, center, and left-wing models. The examples illustrate how the semantic framing of Bundestag differs across ideological contexts.

6.2 Recovering Schwartz's Value Structure in Embedding Space

Having examined the embedding space on its own and analyzed how the key terms differ in both their overall semantic stability and their concrete neighborhood structures, we now extend the analysis by incorporating embeddings of Schwartz's values. Up to this point, semantic variations have been assessed without an external point of reference, focusing on how political terms relate to other words within the embedding space. The following analyses instead introduces Schwartz values as interpretive anchors that allow us to assess whether observed semantic differences align with psychological observations associated with political ideology.
At this stage, the goal is no longer to make semantic change visible through neighborhood structures or spatial separation, but to examine how closely political terms are associated with specific values, and how these associations vary across ideological embedding models. In doing so, semantic differences are translated into value-based alignments that can be compared systematically across left-wing, centrist, and right-wing discourse. As a first step, we take advantage of the fact that Schwartz's theory organizes values in a circular structure defined by compatibility and tension between value types. We therefore examine whether this primarily psychological relational structure can be recovered within the language-based embedding space. For this purpose, we focus on the four higher-level value groups rather than on individual values, as these groups capture the central oppositions of the theory more directly and reduce noise introduced by fine-grained distinctions. The higher-level values are constructed by averaging the embedding vectors of the individual value labels belonging to each group, with "Offenheit für Wandel" (Openness to Change) comprising "Selbstbestimmung" (self-direction), "Stimulation" (stimulation), and "Hedonismus" (hedonism), "Selbsttranszendenz" (Self-Transcendence) comprising "Universalismus" (universalism) and "Wohlwollen" (benevolence), "Bewahrung" (Conservation) comprising "Tradition" (tradition), "Konformität" (conformity), and "Sicherheit" (security), and "Selbststeigerung" (Self-Enhancement) comprising "Macht" (power) and "Leistung" (achievement) (see figure 4.1).
We conduct this analysis for all three value operationalizations (word based, sentence based, axis based) used in this study, providing a baseline for evaluating how well the embeddings capture the theoretical organization of values before relating them to political key terms.

6.2.1 Word based Embeddings

Table 6.9 examines to what extent the word-based value embeddings reproduce the relational logic of Schwartz's higher-level value circle. The theoretical expectation is that the two opposed pairs - "Openness to Change" vs. "Conservation" and "Self-Transcendence" vs. "Self-Enhancement" - should display lower cosine similarity than non-opposed pairs, as they represent motivational conflicts within the theory.
The results, however, present a mixed picture. While "Self-Transcendence" vs. "Self-Enhancement" shows the lowest similarity in the table (0.382), consistent with theoretical expectations for an opposed value pair, the second theoretically opposed pairing, "Open-

Higher-level Value A	Higher-level Value B	Cosine Similarity	Theoretical Relation (acc. to Schwartz)
Openness to Change	Self-Transcendence	0.556	Adjacent
Openness to Change	Conservation	0.507	Opposed
Openness to Change	Self-Enhancement	0.453	Adjacent
Self-Transcendence	Conservation	0.479	Adjacent
Self-Transcendence	Self-Enhancement	0.382	Opposed
Conservation	Self-Enhancement	0.435	Adjacent

Table 6.9: Pairwise cosine similarities between averaged **word** embedding vectors for Schwartz higher-level value groups ("Openness to Change", "Self-Transcendence", "Conservation", "Self-Enhancement"). The column Theoretical Relation (acc. to Schwartz) indicates whether a value pair is theoretically adjacent or opposed in the Schwartz value circle. Adjacent value pairs are expected to exhibit higher cosine similarity than opposed value pairs.

ness to Change" vs. "Conservation" (0.507), does not stand out as particularly dissimilar. In fact, its cosine similarity exceeds that of several non-opposed pairings, such as "Conservation" vs. "Self-Enhancement" (0.435) or "Openness to Change" vs. "Self-Enhancement" (0.453), and is only slightly lower than the similarity observed between the adjacent values "Self-Transcendence" and "Conservation" (0.479). This overlap indicates that, at least for some higher-level value dimensions, theoretically opposed values are not consistently separated from adjacent ones in the word embedding space.
At the same time, a clear-cut opposition between value representations is not to be expected in a distributional semantic model as all values are represented as words that share basic linguistic properties - most notably their function as abstract nouns - and therefore necessarily exhibit a baseline level of semantic similarity. As a result, complete separation between opposed values (cosine similarity of -1) is not realistic in an unmanipulated word-based embedding space. Taken together, the results suggest that the value embeddings provide an informative but imperfect approximation of Schwartz's higher-level structure, which can nonetheless serve as a useful reference point for the subsequent analyses.

6.2.2 Sentence based Embeddings

Table 6.10 reports pairwise cosine similarities between the higher-level value groups when values are represented using sentence-based embeddings. As in the word-based analysis, the aim is to assess whether the relational structure proposed by Schwartz, particularly the opposition between "Openness to Change" and "Conservation", and between "Self-Transcendence" and "Self-Enhancement" is reflected in the embedding space. In contrast to the previous table 6.9, all cosine similarities are negative, indicating that the sentence-based value representations are generally positioned further apart and capture stronger semantic differentiation between value groups.
The pattern with respect to the theoretical oppositions is again mixed. The pair "Openness

Higher-level Value A	Higher-level Value B	Cosine Similarity	Theoretical Relation (acc. to Schwartz)
Openness to Change	Self-Transcendence	-0.341	Adjacent
Openness to Change	Conservation	-0.453	Opposed
Openness to Change	Self-Enhancement	-0.506	Adjacent
Self-Transcendence	Conservation	-0.210	Adjacent
Self-Transcendence	Self-Enhancement	-0.150	Opposed
Conservation	Self-Enhancement	-0.272	Adjacent

Table 6.10: Pairwise cosine similarities between averaged **sentence** embedding vectors for Schwartz higher-level value groups, using SIF-corrected sentence embeddings. The column Theoretical Relation (acc. to Schwartz) indicates whether a value pair is theoretically adjacent or opposed in the Schwartz value circle. Adjacent value pairs are expected to exhibit higher cosine similarity than opposed value pairs.

to change"-"Conservation" shows one of the lowest similarities (-0.453), which is consistent with their conceptual opposition in Schwartz's model. By contrast, "Self-Transcendence" and "Self-Enhancement", while theoretically opposed, does not exhibit a particularly low similarity compared to the other non-opposed pairs (-0,150). At the same time, some non-opposed combinations such as "Openness to change" and "Self-enhancement" display even stronger negative similarity (-0.506).
Overall, the sentence-based results suggest a sharper separation between value groups than observed in the word-based embeddings, but not a cleaner reproduction of the theoretical circular structure. Instead of consistently reflecting the expected oppositions, the sentence embeddings capture broader differences in how values are described and contextualized in language. As such, these embeddings provide a complementary but not strictly more faithful approximation of Schwartz's higher-level value structure, reinforcing the need to interpret future value–term alignments with appropriate caution in the following analyses.

6.2.3 Axis based Embeddings

Lastly table 6.11 reports the cosine similarities between higher-level value groups when values are represented as semantic axes rather than as words or sentences. In this approach, as mentioned, each value and therefore each value group is modeled by averaging multiple contrastive directions in the embedding space, constructed from opposing anchor pairs, which emphasizes opposition and relative positioning rather than shared lexical or definitional content. As a result, cosine similarities primarily reflect whether two value axes point in similar or opposing directions.
The results show again a mixed correspondence with Schwartz's theoretical structure. The opposition between "Openness to Change" and "Conservation" is reflected in a comparatively low similarity (0.030), consistent with their placement on opposite sides of the value circle. By contrast, the theoretically opposed pair "Self-Transcendence" and "Self-Enhancement" (0.208) does not stand out as particularly dissimilar relative to several

Higher-level Value A	Higher-level Value B	Cosine Similarity	Theoretical Relation (acc. to Schwartz)
Openness to Change	Self-Transcendence	0.174	Adjacent
Openness to Change	Conservation	0.030	Opposed
Openness to Change	Self-Enhancement	0.193	Adjacent
Self-Transcendence	Conservation	0.498	Adjacent
Self-Transcendence	Self-Enhancement	0.208	Opposed
Conservation	Self-Enhancement	0.282	Adjacent

Table 6.11: Pairwise cosine similarities between averaged **axis-based** value embeddings for Schwartz higher-level value groups. The column Theoretical Relation (acc. to Schwartz) indicates whether a value pair is theoretically adjacent or opposed in the Schwartz value circle. Adjacent value pairs are expected to exhibit higher cosine similarity than opposed value pairs.

non-opposed pairings. This indicates that the axis-based representations do not uniformly model the expected motivational conflicts.
At the same time, the generally low similarity values suggest that the axis-based method enforces a stronger separation between value groups overall. This reflects the fact that axes encode values as contrasts rather than as descriptive content, reducing shared semantic overlap by construction. Taken together, the results indicate that axis-based embeddings still only partially reproduces the specific oppositional structure proposed by Schwartz. This makes them a complementary perspective that highlights contrasts more strongly, while remaining subject to similar limitations as the other value operationalizations.

6.3 Semantic Change and Value Associations

Having established that the different value operationalizations provide a workable though not perfectly theory-faithful-approximation of Schwartz's higher-level structure, the analysis now turns to the central question of this thesis: whether the well-documented psychological association between value priorities and political ideology can also be detected at the level of political language. Building on the hypothesis introduced in the theoretical and methodological chapters, the following section examines whether political terms systematically shift their value proximity depending on the ideological embedding model in which they are represented. Concretely, if values such as power/security/tradition/conformity/achievement are more strongly associated with right-wing orientations, we would expect that terms embedded in the right-wing corpus exhibit higher cosine similarity to these values, whereas the same terms embedded in left-wing discourse should move closer to values such as universalism/benevolence/self-direction/stimulation/hedonism. The goal of the upcoming analyses is therefore not merely to show that meanings differ across ideological contexts, but to test whether these differences follow the value–ideology regularities reported in political psychology by observing systematic changes in term–value

alignments across the three embedding spaces.

6.3.1 Word based Value-Term Associations

Policy Terms

The word-based analysis provides us with a first test of whether value–ideology associations known from political psychology can be recovered with the distributional proximity between political terms and value labels. Tables 6.12 and 6.13 summarize these results in two complementary ways. Table 6.12 shows, for each value, the policy term whose cosine similarity to that value varies most strongly across the left-, center-, and right-wing embedding models. In other words, the table highlights the term–value pairs that are most sensitive to ideological differences. The figure should therefore be read as a set of "most sensitive" examples rather than a representative overview of all term–value relations. Concretely, each row is interpreted by first identifying the value–term pair, then noting the expected ideology based on Schwartz's theoretical expectations, and finally checking which embedding model (Left, Center, or Right) yields the highest cosine similarity. If the hypothesized pattern held strongly, left-associated values would most often peak in the left-wing model and right-associated values in the right-wing model. Figure 6.13 provides a directional summary of these associations by evaluating whether value–policy term proximity shifts in the theoretically expected ideological direction. Rather than identifying which of the three models yields the single highest cosine similarity, this measure simply tests whether similarity increases toward the expected ideological pole - that is, whether left-associated values are closer to policy terms in the left-wing model than in the right-wing model, and analogously for right-associated values. The center model is deliberately ignored in this evaluation, as the aim is not to assess ideological centrality but to capture directional alignment along the left–right axis.

When read in this way, the word-based results appear mixed and, in several instances, counterintuitive. Although some rows align with the expected direction (e.g., cases in which a left-associated value shows its highest similarity in the left-wing model or a right-associated value peaks on the right), a substantial number of the most divergent pairings do not follow this logic for example the term-value pairs "Sicherheit"-"Umweltschutz" (security-environmental protection) or "Tradition"-"Migration" (tradition-migration) peak in the left-wing model even though the theoretical expection would be the right-wing model. Also the center model frequently shows the highest similarity even for values that are theoretically linked to one ideological pole, suggesting that the observed variation is not simply a matter of left- versus right-specific value anchoring. Moreover, several highly ranked rows are driven by relatively modest shifts in absolute similarity, indicating that "high range" here captures dispersion across models but does not necessarily imply a clear ideological re-positioning toward the expected pole. This is consistent with the general limitation of word-based value representations: value labels such as universalism or self-direction are abstract theoretical terms, and their lexical usage in parliamentary discourse may be sparse, context-dependent, or expressed through paraphrase rather than

Value	Policy Term	Left	Center	Right	Expected Ideology	Range
wohlwollen	umweltschutz	20.7	21.6	37.1	Left	16.4
selbstbestimmung	energiewende	28.8	14.4	21.5	Left	14.5
universalismus	migration	21.3	30.3	16.0	Left	14.3
stimulation	energiewende	28.9	21.0	14.8	Left	14.0
hedonismus	integration	13.5	17.1	23.2	Left	9.7
konformität	integration	29.1	25.7	15.8	Right	13.3
macht	migration	26.2	20.8	14.3	Right	11.9
sicherheit	umweltschutz	39.1	27.3	37.0	Right	11.8
tradition	migration	19.3	14.7	7.8	Right	11.4
leistung	energiewende	22.6	13.3	14.5	Right	9.2

Table 6.12: Top ten policy term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models. Cosine similarities are scaled for readability; range and variance capture cross-model dispersion, and higher ranks indicate stronger ideological sensitivity.

Expected Ideology	N	Directional Match Rate
Left	50	66.0
Right	50	34.0

Table 6.13: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

direct naming.
Table 6.13 makes this limitation more explicit by summarizing alignment in terms of a directional match rate. Even under a directional criterion that does not treat the center as a competing ideological endpoint, the word-based patterns remain far from decisive, especially for right-associated values.

Institutional Terms

Tables 6.14 and 6.15 both report results for institutional terms and should be read in the same way as the previous word-based table: for each value, the tables display the single institutional term–value pairing with the highest ideological range in cosine similarity across the left-, center-, and right-wing embedding models.

Across both figures, the results remain mixed and do not consistently reproduce the expected value–ideology correspondences. While some pairings show the anticipated pattern, such as right-associated values peaking in the right-wing model for specific institutional terms, many of the strongest range cases do not align with expectations. In particular, the center model frequently yields the highest cosine similarity even for values that are theoretically associated with one ideological pole, which suggests that

Value	Institutional Term	Left	Center	Right	Expected Ideology	Range
selbstbestimmung	grundgesetz	31.4	25.5	41.9	Left	16.4
stimulation	föderalismus	21.9	8.4	15.5	Left	13.6
hedonismus	rechtsstaat	35.4	23.9	23.1	Left	12.3
universalismus	rechtsstaat	30.4	28.9	21.8	Left	8.6
wohlwollen	bundesregierung	26.9	28.7	23.1	Left	5.5
tradition	grundgesetz	19.7	7.9	23.5	Right	15.6
sicherheit	sozialstaat	29.1	18.2	31.2	Right	12.9
leistung	sozialstaat	12.6	25.3	16.9	Right	12.7
konformität	rechtsstaat	26.9	14.4	21.5	Right	12.4
macht	sozialstaat	27.3	31.4	24.6	Right	6.8

Table 6.14: Top ten institutional term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models. Cosine similarities are scaled for readability; range and variance capture cross-model dispersion, and higher ranks indicate stronger ideological sensitivity.

Expected Ideology	N	Directional Match Rate
Left	50	60.0
Right	50	38.0

Table 6.15: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

dispersion across models is not equivalent to clear ideological re-positioning. This is especially plausible for institutional concepts, where much of the vocabulary is shaped by legal or procedural language that is shared across parties and tends to be used in relatively standardized ways.
Table 6.15 summarizes these patterns. Although left-associated values show a majority of matches under this measure, alignment remains substantially weaker for right-associated values, reinforcing the overall picture of uneven and only partial value–ideology correspondence in the word-based institutional analysis.

Cultural Terms

Table 6.16 and 6.17 shift the focus to cultural terms, which differ from policy and institutional language in that they are less constrained by formal usage and more directly tied to identity, belonging, and symbolic meaning. As in the previous analyses, Table 6.16 reports, for each value, the cultural term–value pairing with the highest ideological range, highlighting those cases in which value proximity varies most strongly across the three embedding models.

Compared to policy and institutional terms, the cultural results display larger ranges and more pronounced shifts, indicating that cultural concepts are more sensitive to ideological

Value	Cultural Term	Left	Center	Right	Expected Ideology	Range
wohlwollen	patriotismus	27.6	23.9	43.3	Left	19.4
stimulation	heimat	19.4	15.9	3.2	Left	16.2
selbstbestimmung	patriotismus	43.9	34.7	48.2	Left	13.5
universalismus	nation	24.3	23.4	36.6	Left	13.2
hedonismus	heimat	9.1	19.8	15.8	Left	10.7
konformität	patriotismus	22.7	17.6	34.6	Right	17.0
macht	patriotismus	33.6	29.4	45.2	Right	15.8
tradition	identität	12.6	15.6	27.5	Right	14.9
sicherheit	identität	22.5	21.7	36.5	Right	14.9
leistung	heimat	9.1	17.2	12.6	Right	8.1

Table 6.16: Top ten cultural term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models. Cosine similarities are scaled for readability; range and variance capture cross-model dispersion, and higher ranks indicate stronger ideological sensitivity.

Expected Ideology	N	Directional Match Rate
Left	50	42.0
Right	50	70.0

Table 6.17: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

framing at the lexical level, which is supported by the data from table 6.1. Terms such as “Patriotismus” (patriotism), “Heimat” (homeland), “Identität” (identity), and “Nation” (nation) recur across multiple rows and are associated with both left- and right-linked values, underscoring their semantic contestedness. Importantly, however, high divergence does not translate into consistent ideological alignment. Several left-associated values, including “Wohlwollen” (benevolence), “Universalismus” (universalism), and “Selbstbestimmung” (self-direction), show their highest similarity in the right-wing embedding for specific cultural terms, while some right-associated values also fail to peak unambiguously in the right-wing model. These patterns suggest that cultural terms function as polyvalent symbols whose value associations shift depending on context, rather than as stable carriers of a single ideological value profile.

Table 6.17 summarizes these observations using the directional match criterion. In contrast to the institutional analysis, right-associated values exhibit a substantially higher directional match rate (70.0%), whereas alignment for left-associated values remains comparatively weak (42.0%). This asymmetry indicates that, within cultural discourse, values linked to order, conformity, and authority are more consistently expressed through culturally loaded terms, while humanitarian or autonomy-oriented values appear to be articulated in more diffuse or indirect ways. At the same time, the overall pattern remains uneven, reinforcing the conclusion that even in the culturally most ideologically charged

domain, word-based value representations capture value–ideology relationships only partially.

6.3.2 Sentence based Value-Term Associations

Policy Terms

In table 6.18 the sentence-based analysis revisits in the policy-terms using embeddings for the values that should incorporate broader contextual informations rather than isolated value labels. As in the previous tables table 6.18 reports, for each value, the policy term that exhibits the largest ideological range in cosine similarity across the left-, center-, and right-wing embedding models for that value.

Value	Policy Term	Left	Center	Right	Expected Ideology	Range
universalismus	energiewende	9.3	16.7	-2.2	Left	18.9
stimulation	umweltschutz	6.4	9.0	-5.4	Left	14.4
selbstbestimmung	migration	4.2	1.6	-8.6	Left	12.9
wohlwollen	asyl	-0.0	12.7	9.6	Left	12.7
hedonismus	integration	-11.3	-7.5	1.4	Left	12.7
leistung	integration	17.1	11.1	-2.0	Right	19.1
tradition	klimaschutz	-5.0	-11.1	5.5	Right	16.6
sicherheit	asyl	-2.2	-14.3	1.8	Right	16.1
macht	inflation	-2.2	-0.4	10.3	Right	12.6
konformität	inflation	7.0	5.5	-3.0	Right	10.1

Table 6.18: Top ten policy term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

Compared to the word-based results, the sentence-based patterns appear somewhat more structured, but still far from uniformly theory-consistent. Several left-associated values, such as “Universalismus” (universalism), “Stimulation” (stimulation), and “Selbstbestimmung” (self-direction), show their highest similarity in the left-wing embedding and decline toward the right, which is in line with theoretical expectations. At the same time, other left-associated values such as “Wohlwollen” (benevolence) or “Hedonismus” (hedonism) peak in the center or even in the right-wing model, indicating that contextualized value descriptions do not automatically yield clearer ideological separation. A similar pattern holds for right-associated values: while some, such as “Tradition” (tradition) or “Macht” (power), show increased proximity in the right-wing embedding, others display weaker or reversed gradients. Overall, the sentence-based table suggests that incorporating context sharpens ideological contrasts in some cases, but does not eliminate overlap or ambiguity across ideological spaces.

Expected Ideology	N	Directional Match Rate
Left	50	44.0
Right	50	54.0

Table 6.19: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

The directional match rates reported in the table 6.19 reinforce this mixed picture. When alignment is evaluated directionally rather than categorically, right-associated values show a modestly higher match rate (54%) than left-associated values (44%), while both remain close to chance-level rather than indicating strong systematic alignment. This contrasts with the word-based policy-term analysis (table 6.13), where left-associated values showed comparatively stronger directional consistency. The reversal suggests that sentence-based representations may capture different aspects of value expression, emphasizing how terms are framed in context rather than how they are lexically named.

Institutional Terms

Tables 6.20 and 6.21 extend the sentence-based analysis to institutional terms, applying the same selection and evaluation logic used for policy terms.

Value	Institutional Term	Left	Center	Right	Expected Ideology	Range
selbstbestimmung	opposition	13.9	-1.2	2.9	Left	15.1
stimulation	opposition	-5.7	-16.4	-20.8	Left	15.0
hedonismus	föderalismus	10.6	-3.6	0.6	Left	14.2
universalismus	opposition	3.2	12.3	-0.4	Left	12.8
wohlwollen	opposition	-8.7	3.0	1.2	Left	11.8
tradition	sozialstaat	-8.5	-1.0	8.1	Right	16.5
konformität	sozialstaat	-3.1	8.9	-5.8	Right	14.7
leistung	opposition	-7.0	-1.9	6.9	Right	13.8
macht	rechtsstaat	3.2	-0.2	12.6	Right	12.8
sicherheit	rechtsstaat	-3.6	-8.3	2.0	Right	10.2

Table 6.20: Top ten institutional term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

The institutional results again reveal substantial variation, but little systematic alignment with theoretical expectations. A striking feature of table 6.20 is the frequent appearance of the term “Opposition” (opposition), which dominates several of the highest-range pairings across both left- and right-associated values like “Stimulation” (stimulation), “Selbstbestimmung” (self-direction) or “Leistung” (achievement). This concentration suggests that

Expected Ideology	N	Directional Match Rate
Left	50	48.0
Right	50	48.0

Table 6.21: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

ideological divergence in institutional language is often driven by the semantic flexibility of a small number of highly salient terms. However, even when divergence is pronounced, the direction of the shift is inconsistent: left-associated values such as "Selbstbestimmung" (self-direction) or "Universalismus" (universalism) do not reliably peak in the left-wing embedding, while right-associated values such as "Tradition" (tradition) or "Macht" (power) do not uniformly show their strongest association in the right-wing model. As in the word-based analysis, the center embedding most of the time occupies an intermediate or even dominant position.
Figure 6.21 summarizes these patterns using the directional match criterion. The symmetry of the Directional Match Rate indicates that, in the institutional domain, sentence-based value representations do not systematically capture ideological shifts. Taken together, table 6.20 and 6.21 suggest that institutional language exhibits limited and unstable value–ideology alignment. The sentence based analysis emphasizes the results from the word based analysis: institutional terms largely operate within procedural frames that limit ideological value expression, making institutional discourse less receptive to value-based semantic differentiation.

Cultural Terms

We now complete the sentence-based analysis by again turning to cultural terms, which previous data has shown to be typically less stable and more directly tied to questions of values.

The resulting patterns in Tables 6.22 again point to substantial dispersion, but only limited theoretical alignment. The terms "Identität" (identity), "Patriotismus" (patriotism), "Leitkultur" (guiding culture), "Heimat" (homeland), and "Nation" (nation) dominate the table and recur across both left- and right-associated values, underscoring their role as semantically contested symbols. However several left-associated values, including "Universalismus" (universalism) and "Stimulation" (stimulation), display their highest similarity in the right-wing or center embeddings, while some right-associated values fail to peak clearly in the right-wing model like "Konformität" (conformity). These results suggest that, even when values are represented through contextualized sentence embeddings, cultural terms do not map cleanly onto the value–ideology structure proposed by Schwartz. Instead, high range values often reflect strong re-framing of culturally loaded concepts rather than systematic movement toward the theoretically expected ideological pole.

Value	Cultural Term	Left	Center	Right	Expected Ideology	Range
universalismus	identität	-11.6	0.1	7.7	Left	19.3
stimulation	solidarität	-11.2	-18.7	-0.5	Left	18.2
selbstbestimmung	patriotismus	15.5	11.1	1.1	Left	14.4
wohlwollen	leitkultur	-16.7	-7.5	-5.3	Left	11.4
hedonismus	identität	0.1	-0.6	-8.5	Left	8.6
tradition	heimat	-4.5	3.9	12.3	Right	16.8
leistung	patriotismus	7.3	-6.8	-8.7	Right	16.0
sicherheit	patriotismus	-11.9	0.5	3.2	Right	15.1
macht	nation	-2.7	6.7	-6.3	Right	13.0
konformität	leitkultur	3.1	-6.7	0.9	Right	9.9

Table 6.22: Top ten cultural term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

Expected Ideology	N	Directional Match Rate
Left	50	52.0
Right	50	44.0

Table 6.23: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

Table 6.23 summarizes these findings using the directional match criterion. In contrast to the word-based cultural analysis (table 6.17), the sentence-based results show only modest alignment for both ideological groups, with match rates close to parity and only slightly above chance. Especially the correct matching rates for the right-wing models dropped significantly (52.0% compared to previous 72%).
Taken together, table 6.22 and 6.23 indicate that the sentence-based embeddings analysis do not consistently align with the theorized value–ideology patterns. Cultural terms remain difficult to anchor to stable value orientations, even though the sentence based embeddings can be understood as richer contextual representations than the simply word based embeddings.

6.3.3 Axis based Value-Term Associations

We now turn to the final embedding format used in this study: the axis-based representation. In this approach, values are not modeled as lexical items like words or sentence-level descriptions, but as directional dimensions in the embedding space, constructed from opposing sets of anchor terms. This representation is intended to capture the term value alignments in a more abstract and relational manner.

Policy Terms

Value	Policy Term	Left	Center	Right	Expected Ideology	Range
stimulation	umweltschutz	9.9	15.9	-1.0	Left	16.8
universalismus	asyl	11.0	-5.3	3.5	Left	16.3
selbstbestimmung	inflation	-10.7	-10.8	3.3	Left	14.2
wohlwollen	inflation	-17.2	-10.6	-4.8	Left	12.4
hedonismus	asyl	1.7	-4.8	4.3	Left	9.1
konformität	inflation	-9.2	-5.3	5.3	Right	14.5
sicherheit	energiewende	13.6	9.9	-0.6	Right	14.1
tradition	klimaschutz	-2.5	-2.2	10.7	Right	13.2
leistung	frieden	-3.1	9.2	-3.3	Right	12.5
macht	sicherheit	-4.3	5.1	6.8	Right	11.1

Table 6.24: Top ten policy term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

Expected Ideology	N	Directional Match Rate
Left	50	52.0
Right	50	30.0

Table 6.25: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

Viewed in comparison to the previous representations, the axis-based policy-term results resemble the word-based patterns more closely than the sentence-based ones. In particular, the directional match rates in table 6.25 again display a clear asymmetry: left-associated values achieve a modest majority of directional matches (52%), whereas right-associated values show substantially weaker alignment (30%). This mirrors the imbalance observed in the word-based analysis and contrasts with the more balanced alignment found for sentence-based embeddings.

At the level of individual high-range pairings seen in table 6.24, this similarity is also reflected in the instability of right-linked value alignment. While some right-associated values like “Tradition” (tradtion) or “Macht” (power) peak in the right-wing embedding for specific policy terms, these cases remain inconsistent, and several left-associated values likewise fail to exhibit a clear leftward shift like “Selbstbestimmung” (self-direction). As in the word-based analysis, high ideological range primarily indicates sensitivity to ideological context rather than reliable convergence toward the expected ideological pole. Taken together, the axis-based results suggest that modeling values as directional oppositions unfortunately does not overcome the asymmetries already present in simpler lexical representations.

Institutional Terms

The axis-based analysis of the institutional terms largely reproduces the patterns already observed in the word- and sentence-based representations. As in the previous cases, table 6.26 shows that the highest-range pairings are concentrated among a small set of institutional terms, most notably "Föderalismus" (federalism) and "Rechtsstaat" (rule of law). These terms recur across values with different expected ideological orientations, indicating that ideological divergence in the institutional domain is driven by the semantic flexibility of a limited number of core concepts rather than by stable value-specific alignment.

Value	Institutional Term	Left	Center	Right	Expected Ideology	Range
universalismus	föderalismus	7.7	-4.0	11.2	Left	15.2
stimulation	rechtsstaat	3.9	0.9	-10.8	Left	14.7
hedonismus	föderalismus	8.8	10.5	-3.3	Left	13.8
selbstbestimmung	föderalismus	0.2	0.0	13.7	Left	13.6
wohlwollen	föderalismus	-14.8	-8.2	-1.4	Left	13.5
macht	föderalismus	-12.4	4.3	-2.1	Right	16.7
konformität	verfassung	-4.4	-9.9	4.2	Right	14.1
sicherheit	rechtsstaat	7.3	19.4	6.4	Right	13.0
leistung	rechtsstaat	-0.1	4.7	-6.7	Right	11.4
tradition	rechtsstaat	-2.7	0.3	-9.4	Right	9.7

Table 6.26: Top ten Institutional term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

Expected Ideology	N	Directional Match Rate
Left	50	44.0
Right	50	58.0

Table 6.27: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

The directional match rates in table 6.27 indicate a modest asymmetry between left- and right-associated values, with right-associated values showing a higher rate of directional alignment (58%) than left-associated ones (44%). While this difference points to somewhat clearer rightward value structuring in institutional language, the overall levels remain moderate and comparable to those observed in the sentence-based analysis. Thus, even when values are modeled as explicit semantic axes, institutional discourse does not exhibit a strong or consistent correspondence with the value–ideology regularities.

Cultural Terms

Finally, we turn to the axis-based representation of cultural terms, completing the comparison across all three value operationalizations used in this study.
In comparison to the other embedding formats, the axis-based analysis of cultural terms occupies an intermediate position. Relative to the word-based results in table 6.17 , in table 6.29 axis-based embeddings substantially reduce the strong right-leaning asymmetry: while the word-based cultural analysis favored right-associated values (70% directional match) over left-associated ones (42%), the axis-based representation reverses this imbalance, yielding a higher match rate for left-associated values (64%) and a neutral outcome for right-associated values (50%). Compared to the sentence-based embeddings (table 6.23), which showed near-chance alignment for both sides (52% left, 44% right), the axis-based approach therefore provides a modest improvement in capturing theoretically expected value–ideology relations for cultural discourse.

Value	Cultural Term	Left	Center	Right	Expected Ideology	Range
selbstbestimmung	identität	12.4	-2.0	14.4	Left	16.4
hedonismus	patriotismus	17.7	4.2	1.6	Left	16.1
stimulation	identität	7.7	-1.1	-8.3	Left	16.0
universalismus	identität	1.4	2.2	12.2	Left	10.7
wohlwollen	heimat	2.0	7.6	9.8	Left	7.9
leistung	nation	3.9	-5.4	-11.8	Right	15.7
konformität	nation	-1.3	-14.4	-3.7	Right	13.1
sicherheit	leitkultur	2.3	14.6	9.0	Right	12.3
tradition	identität	-4.1	-1.9	8.0	Right	12.1
macht	heimat	2.2	-6.0	1.0	Right	8.2

Table 6.28: Top ten Cultural term–value associations with the highest ideological divergence across left-, center-, and right-wing embedding models using sentence-based value representations.

Expected Ideology	N	Directional Match Rate
Left	50	64.0
Right	50	50.0

Table 6.29: Directional match rates, indicating whether value proximity increases toward the theoretically expected ideological pole.

At the same time, table 6.28 makes clear that this improvement is selective rather than general. High-range associations continue to cluster around a small set of culturally contested symbols such as “Identität” (identity), “Patriotismus” (patriotism), “Heimat”

(homeland), and “Nation” (nation), which appear across multiple values with opposing ideological expectations. This pattern persists across all three embedding methods and indicates that ideological variation in cultural language is driven less by stable value anchoring than by the symbolic flexibility of a few central terms. Thus, while axis-based embeddings sharpen directional structure compared to sentence-based representations they do not fully resolve the instability of value alignment observable with the cultural terms.

7 Discussion

Now having presented the empirical results of our analyses we can turn in the following section to their discussion and broader interpretation. Rather than reiterating the descriptive findings, the discussion tries to situate the observed patterns of semantic variation within the theoretical framework of semantic change, ideological framing, and value-based meaning construction, and reflects on their implications for the study of political language.

7.1 Overall Semantic Change in Political Terms

First we will talk about the overall observed semantic change in the key political terms in Chapter 6. Across the full set of predefined key terms, stability scores remain relatively low in absolute terms, with the highest score still below 0.4 and several terms falling close to the lower bound. Institutional terms show the strongest stability on average, driven in particular by constitutional references such as "Grundgesetz" (basic law) and "Demokratie" (democracy), whose comparatively high scores suggest that they retain similar semantic positions across ideological contexts. Yet at the same time, the institutional category also contains more contested terms (e.g., "Bundesregierung" (federal government), "Föderalismus" (federalism), "Sozialstaat" (welfare state)), highlighting that institutional language is not uniformly stable. Policy terms occupy an intermediate position: some terms exhibit comparatively high stability (notably "Inflation" (inflation) and "Migration" (migration)), while others such as "Gerechtigkeit" (justice), "Energiewende" (energy transition), and "Integration" (integration) show markedly lower stability scores. Cultural terms display the lowest aggregate stability, largely driven by strongly contested symbols, most notably "Leitkultur" (guiding culture), which shows near-total divergence, as well as "Nation" (nation) or "Patriotismus" (patriotism). Therefore, we can confirm our hypothesis that cultural terms are the most sensible for semantic change.
While the results point to substantial semantic variation across the examined political vocabulary, the reported stability scores must be interpreted with considerable methodological caution. In particular, the two components used to calculate the stability score, cosine similarity and neighbor difference, prove to be insufficient to provide a definitive assessment of semantic change. Cosine similarity captures shifts in the global position of a political term within the embedding space; however, such positional change in itself has no intrinsic semantic meaning. A political term may remain globally stable while undergoing substantial semantic change in its contextual usage, or conversely, shift position without reflecting a meaningful change in meaning. Neighbor difference, by contrast, is more closely related to semantic change, as it captures alterations in local contextual associations. Yet the consistently high neighbor difference observed across most political terms limits its interpretability. Because neighbor difference values are so high across the board, they contribute only weakly to the overall stability score. As a result, cosine similarity is implicitly overemphasized in the calculation, despite its limited semantic interpretability. The resulting stability score therefore risks overstating semantic stability across embedding spaces. This insight is particularly relevant given that many studies on semantic change rely on similar metric combinations. Future research should address this limitation by contextualizing neighbor difference values relative to average neighbor difference across

the entire vocabulary, thereby increasing their interpretive value for assessing semantic change in political terms. More generally, this highlights that measures of semantic change are never theory-neutral. The choice and combination of metrics implicitly define what is understood as semantic stability or change, often without being made explicit. In practice, this means that methodological decisions can shape substantive conclusions about meaning change as much as the underlying data. Making these assumptions visible and empirically testable is therefore a crucial step for advancing methodological rigor in semantic change research.
Another interpretation of the consistently high neighbor divergence observed across the many political terms is that it reflects not primarily semantic change, but substantial differences in usage and vocabulary across ideological corpora. If political actors draw on largely distinct lexical repertoires, the overlap at the level of local semantic neighborhoods will necessarily be limited, even when reference to the same political terms exists. From this perspective, ideological differentiation may operate less through systematic semantic change of shared terms and more through selective vocabulary choice and differential topic articulation. This has important implications for political language research, suggesting that semantic change may not always be the most informative lens for capturing ideological differences. Instead, greater analytical value may lie in identifying the key political terms around which ideological discourse is organized, rather than in tracking semantic change within a limited set of shared vocabulary. What is the core terminology of an ideology? Future research should therefore place greater emphasis on vocabulary divergence, term selection, and the broader lexical composition of ideological discourse.

7.2 The Limits of Mapping Psychological Patterns onto Distributional Semantics

Regarding the traceability of psychological value–ideology patterns in language-based representations, we can state the following:
Across the value-based analyses, the expected mapping from Schwartz's established value–ideology correlations onto the embedding spaces is only reproduced in a limited and uneven way. While term–value proximities clearly vary across left, center, and right models, the direction of these differences often fails to match the theoretically expected ideological pole. This mixed pattern occurs across all our three types of terms and shows that cosine similarities between value representations and semantic domains vary too strong to form a stable or coherent ideological value cluster in the embedding spaces.
Comparing word-based, sentence-based, and axis-based value operationalizations, there is no consistent indication that one representation yields systematically clearer value–ideology alignment than the others. Instead, the results across formats repeatedly converge on the same qualitative conclusion: substantial dispersion in term–value associations exists, but it does not reliably translate into the expected ideological gradients. Even where contextualization or abstraction is increased (sentence- and axis-based approaches),

directional match rates remain close to chance levels in several settings, and the center embedding frequently remains intermediate or dominant rather than supporting a clean polarization pattern.
A plausible interpretation of the weak alignment is that the theoretical level at which Schwartz's values operate simply does not translate straightforwardly into distributional patterns in parliamentary speech. Values are defined as highly abstract motivational goals, whereas embeddings capture associations arising from situated linguistic usage; this gap makes it difficult for value structure to appear as stable geometric separation in a Word2Vec space. This tension was already apparent when testing whether the higher-level circumplex logic can be recovered in the value embeddings themselves.
Throughout the value–correlation analyses, the center embedding repeatedly complicates a simple left-versus-right interpretation. In multiple domains and operationalizations, the center model often occupies as one would expect an intermediate position, but it can also appear as the highest similarity point even for values theoretically associated with a specific ideological pole. This pattern suggests that cross-model dispersion is not equivalent to a clean ideological re-positioning and that the center space does not behave like a mere midpoint between two extremes.
Taken together, the findings suggest that mirroring psychological value–ideology patterns in language-based vector spaces may only be feasible under more restricted circumstances. One limiting factor concerns the nature of the training data: embeddings trained on highly institutionalized and strategically constrained language, such as parliamentary speeches, may lack the linguistic variability required for abstract psychological value structures to emerge in a stable form. Training data drawn from more naturalistic and less role-bound contexts, in which value expressions are articulated more directly and with greater lexical diversity, may be better suited to capturing such structures in distributional semantics.
This leads to implication concerning research design. Rather than assuming a direct one-to-one translation from psychological value structures to distributional representations of language, the results suggest an incremental approach in which researchers first identify those values and value embeddings that show stable and theoretically plausible patterns in embedding spaces, and then build subsequent analyses on these empirically supported configurations to refine the link between linguistic patterns and psychological theory.

7.3 Revisiting Boutyline's Axis-Based Approach to Conceptual Meaning

This leads us to our discussion of the concept–context dichotomy explained by Boutyline and Arseniev-Koehler (2025). As discussed, one major aspect of working with word embeddings, especially in the social sciences, is the attempt to measure abstract concepts that are difficult to pin down. Word2Vec as an embedding methodology, however, always relies on some form of contextual information. To bridge this gap from context to concept, Boutyline recommends the construction of axis-based embeddings. In this thesis, this approach was implemented to create embeddings for the investigated values.

From a conceptual perspective, Boutyline's recommendation to use axis-based embeddings provides a clear and theoretically well-motivated strategy for approximating abstract concept spaces within distributional models. By explicitly defining semantic dimensions through opposing anchor terms, axis-based approaches aim to reduce the influence of surface-level contextual regularities and to isolate the conceptual feature of interest. In this sense, the approach directly addresses the concept–context tension inherent in word embedding methodologies.

Empirically, however, the results of this thesis suggest that the advantages of axis-based embeddings are limited in the present setting. While axis-based value embeddings do reduce some of the noise associated with single word representations, they do not consistently yield clearer or more theoretically aligned value–ideology patterns than word- or sentence-based approaches. Across domains, the directional match rates and ideological divergences observed with axis-based embeddings remain comparable to those obtained with alternative operationalizations, rather than constituting a clear improvement.

However, one can not take this interpretation too far. One possible interpretation of our axis-based results is that language use simply does not mirror the patterns observed in psychological research by Schwartz. Even if semantic axes successfully represent abstract concepts like values within the embedding space, these linguistic representations may not correspond directly to value–ideology relationships measured through surveys or experiments. In this sense, the lack of alignment between our empirical results and the the expected results does not automatically imply that the axis-based approach failed, but rather suggests that psychological findings may not translate one-to-one into patterns of language use. Moreover, as with all embedding-based methods, axis representations remain fundamentally dependent on the training data. If the underlying corpus does not express values in ways that closely correspond to psychological constructs, as may be the case in highly institutionalized and strategic political discourse, then even well constructed axes will reflect these constraints rather than abstract motivational structures.

Taken together, the findings leave the approach recommended by Boutyline in an ambivalent position. While axis-based embeddings offer a principled and theoretically appealing way to operationalize abstract concepts, their empirical performance in this study does not provide clear evidence that they systematically outperform alternative embedding strategies in capturing value–ideology relationships. At the same time, the observed limitations do not allow for a definitive assessment of whether the approach fails to represent the intended concepts or whether the concepts themselves cannot be reliably mirrored in language-based representations. This suggests that further research is needed that focuses explicitly on evaluating the conceptual strength of embedding-based representations.

7.4 Limits of the Methodology and Data and future Endeavors

Finally we shortly discuss the limitations of the methodological choices and data underlying this study.
A central methodological limitation of this study concerns the use of parliamentary speeches as training data for the word embedding models. What was initially expected to be a strength of the dataset may have had a constraining effect on the analytical insights that could be derived from the results. Political speeches are produced in a highly institutionalized, strategic, and norm-governed setting. Speakers operate under role-specific constraints, adhere to formal conventions, and often pursue persuasive or rhetorical goals rather than expressing values or beliefs directly. As a result, language use in this context may prioritize framing, signaling, and strategic ambiguity over explicit articulation of underlying motivational structures. While this makes parliamentary discourse well suited for analyzing ideological contestation and semantic framing, it may limit its suitability for capturing abstract psychological constructs such as values. Consequently, embeddings trained on this data may reflect patterns of political communication rather than broader value structures as conceptualized in psychological theory.
A further methodological consideration arises from asymmetries between the three ideological training corpora. Although care was taken to construct comparable corpora, differences in corpus size and party composition remain. These asymmetries can influence embedding geometry, as Word2Vec models are sensitive to frequency distributions and lexical variations. Even subtle differences in how often terms are used affect cosine similarities and neighborhood structures without necessarily a change in the semantic. As a result, some observed differences across ideological embedding spaces may reflect corpus-specific properties rather than substantive ideological distinctions. This does not invalidate the findings, but it suggests that comparisons between embeddings should be interpreted with caution, particularly when differences are small or inconsistent across domains.
Finally, the use of static word embeddings introduces structural limitations that are especially relevant for political and sociological analysis. Static embeddings assign a single vector representation to each word, regardless of context, speaker, or communicative intent. This makes it difficult to account for polysemy, strategic ambiguity, and context-dependent meaning shifts that are common in political language. Moreover, static embeddings aggregate usage patterns across an entire corpus, thereby smoothing over situational and discursive variation that may be theoretically important. While static Word2Vec models are well suited for detecting broad distributional differences and long-term semantic patterns, their capacity to capture fine-grained ideological meaning or abstract sociological constructs is inherently limited. The results of this thesis reflect both the strengths and the boundaries of this methodological choice.
Transformer-based and pretrained language models could extend this line of research in several ways. By generating context-sensitive representations, these models make it possible to distinguish between different uses of the same term across topics, speakers, or argumentative settings, rather than collapsing all occurrences into a single vector. This would allow future analyses to examine not only whether semantic positions differ across ideolo-

gies, but also how ideological meaning is constructed within specific discursive contexts. In addition, pretrained models offer the opportunity to decouple representation learning from a single, highly institutionalized corpus, enabling comparisons between parliamentary language and broader, less constrained forms of political communication. Finally, transformer-based approaches could support downstream tasks - such as classification or value prediction-that provide an external criterion for assessing whether abstract concepts are being captured in a way that is analytically useful.

8 Appendix

8.1 Overview of Semantic Axes

Each semantic axis is defined by a set of antonymous word pairs used as anchors for constructing axis-based embeddings. The table below lists all anchor pairs employed in the analysis.

Table 8.1: Anchor word pairs used for constructing semantic axes

Axis	Positive Pole	Negative Pole
Macht	mächtig; stark; stärker; gewaltig; gewaltiger; einflussreich; bedeutsam; herrschend; autoritär; befehlen; kontrollierend; herrschaft; einfluss; prestige; mächtiger; mächtige; mächtigsten; mächtigste; herrscher; könig; anführer	machtlos; schwach; schwächer; harmlos; harmloser; wirkungslos; unbedeutend; unterwürfig; gehorchen; abhängig; gleichheit; einflusslosigkeit; verachtung; machtloser; machtlose; ohnmächtig; diener; bauer; untergebener
Universalismus	universalismus; tolerant; gerecht; inkludierend; weltoffen; solidarisch; nachhaltig; verstehend; wertschätzend; menschenrechte; gleichberechtigung; solidarität; frieden; umweltschutz; universalistisch; universalistischer; universalistische; weltbürger; umweltschützer; gleichheitsbefürworter; pazifist	partikularismus; intolerant; ungerecht; ausschließend; engstirnig; unsolidarisch; zerstörerisch; urteilend; verachtend; diskriminierung; ungleichheit; egoismus; krieg; ausbeutung; partikularistisch; partikularistischer; partikularistische; nationalist; umweltzerstörer; elitist; kriegstreiber
Wohlwollen	wohlwollen; hilfsbereit; loyal; verzeihend; ehrlich; verantwortungsvoll; freundschaftlich; liebevoll; treue; fürsorge; gemeinschaft; vertrauen; kooperation; wohlwollend; wohlwollender; wohlwollende; wohlwollendem; freund; partner; familie	egoismus; selbstsüchtig; illoyal; nachtragend; unehrlich; verantwortungslos; feindselig; gleichgültig; verrat; vernachlässigung; isolation; misstrauen; konkurrenz; egoistisch; egoistischer; egoistische; egoistischem; feind; verräter; fremder
Tradition	tradition; brauchtum; ritual; religion; glaube; bescheidenheit; symbol; kultur; gemeinschaft; ordnung; kontinuität; traditionell; traditioneller; traditionelle; traditionellsten; traditionellem; gläubiger; priester; patriarch	moderne; neuheit; innovation; säkularismus; skepsis; arroganz; beliebigkeit; individualismus; egoismus; anarchie; wandel; modern; moderner; moderne; modernsten; modernem; atheist; rebell; revolutionär

Axis	Positive Pole	Negative Pole
Konformität	konformität; gehorsam; diszipliniert; höflich; respektvoll; ordentlich; ehrfürchtig; pflichtbewusst; respekt; ehrfurcht; konform; konformer; konforme; bürger; funktionär; beamter	nonkonformität; ungehorsam; undiszipliniert; unhöflich; respektlos; unordentlich; verachtend; verantwortungslos; respektlosigkeit; verachtung; nonkonform; nonkonformer; nonkonforme; dissident; oppositioneller; anarchist
Sicherheit	sicherheit; stabil; geordnet; ruhig; harmonisch; gesund; rein; zugehörig; schutz; geborgenheit; vertrauen; kontinuität; solidarität; sicher; sicherer; sichere; sicherem	unsicherheit; instabil; chaotisch; unruhig; konflikthaft; krank; schmutzig; entfremdet; bedrohung; angst; misstrauen; unterbrechung; isolation; unsicher; unsicherer; unsichere; unsicherem
Leistung	leistung; erfolgreich; fähig; kompetent; ehrgeizig; einflussreich; bedeutsam; anerkennung; prestige; aufstieg; ruhm; ehrgeiz; erfolg; leistender; leistende; leistungsfähig; leistungsstark; gewinner; champion; führer	versagen; erfolglos; unfähig; inkompetent; faul; bedeutungslos; unbedeutend; ablehnung; verachtung; abstieg; schande; faulheit; scheitern; versagender; versagende; unfähig; leistungsschwach; verlierer; versager; mitläufer
Hedonismus	hedonismus; lustvoll; genussvoll; vergnügt; lebenslustig; heiter; spielerisch; vergnügen; lust; spaß; sinnlichkeit; hedonistisch; hedonistischer; hedonistische; hedonistischem; feinschmecker; genießer; partygänger; spieler; bonvivant	askese; enthaltsam; verzichtend; stoisch; pflichtbewusst; streng; ernst; verzicht; enthaltung; langweile; abstinenz; asketisch; asketischer; asketische; asketischem; asket; verzichter; einzelgänger; moralist; puritaner
Stimulation	stimulation; aufregend; neugierig; dynamisch; abwechslungsreich; herausforderung; spannung; erregung; stimulierend; stimulierender; stimulierende; entdecker; pionier; abenteurer; reisender	langeweile; öde; desinteressiert; träge; eintönig; bequemlichkeit; ruhe; gleichgültigkeit; langweilig; langweiliger; langweilige; sesshafter; traditionalist; biederer; zuhausebleiber

Axis	Positive Pole	Negative Pole
Selbstbestimmung	selbstbestimmung; unabhängig; frei; autonom; kreativ; neugierig; eigenständig; selbstständig; freiheit; initiative; gestaltung; selbstkontrolle; unabhängigkeit; selbstbestimmt; selbstbestimmter; selbstbestimmte; rebell; pionier; künstler; erfinder; führer	fremdbestimmung; abhängig; unfrei; untergeordnet; angepasst; gleichgültig; fremdgesteuert; unselbstständig; zwang; passivität; gehorsam; fremdkontrolle; abhängigkeit; fremdbestimmt; fremdbestimmter; fremdbestimmte; untertan; folgender; beamter; nachahmer; gehorsamer

8.2 Sentence-Based Embeddings

The following sentences were used to construct sentence-based embeddings for each value dimension. The descriptions are based on motivational definitions of the respective values.

Value	Sentence Description
Universalismus	Das motivationale Ziel von Universalismus besteht in Verständnis, Wertschätzung, Toleranz und Schutz für das Wohlergehen aller Menschen und der Natur. Diese Aspekte ergeben sich aus den Erfordernissen des Überlebens von Gruppen und Individuen, insbesondere im Kontakt mit Mitgliedern einer Outgroup sowie angesichts begrenzter natürlicher Ressourcen.
Wohlwollen	Beim Wert Wohlwollen steht der Erhalt und die Verbesserung des Wohlergehens von Menschen im Mittelpunkt, mit denen man in häufigem persönlichen Kontakt steht. Dieser Wert leitet sich sowohl aus dem universalen Erfordernis positiver sozialer Interaktion als auch aus dem Bedürfnis nach Zugehörigkeit ab.
Tradition	Das motivationale Kernziel des Wertes Tradition besteht in Respekt, Verpflichtung und Akzeptanz gegenüber den Bräuchen und Ideen der eigenen Kultur oder Religion. Traditionen entstehen aus gemeinsamen Erfahrungen von Gruppen und werden als stabilisierende Symbole für kollektive Identität und Kontinuität wertgeschätzt.

Value	Sentence Description
Konformität	Der Wert Konformität fokussiert sich auf die Zurückhaltung von Handlungen, Neigungen und Impulsen, die soziale Normen verletzen oder anderen schaden könnten. Er entspringt der Notwendigkeit, soziale Interaktionen innerhalb von Gruppen reibungslos zu gestalten.
Sicherheit	Im Zentrum des Wertes Sicherheit stehen Unversehrtheit, Harmonie und Stabilität der Gesellschaft, persönlicher Beziehungen und des eigenen Selbst. Dieser Wert leitet sich aus grundlegenden individuellen und kollektiven Schutzbedürfnissen ab.
Macht	Der Wert Macht beschreibt das motivationale Ziel, sozialen Status, Prestige sowie Kontrolle oder Vorherrschaft über Personen und Ressourcen zu erlangen. Er entsteht aus funktionalen Anforderungen sozialer Institutionen und der Organisation von Gruppen.
Leistung	Das definierende Merkmal des Wertes Leistung ist persönlicher Erfolg durch die Demonstration von Kompetenz gemäß sozialer oder kultureller Standards, um soziale Anerkennung zu erhalten und Zugang zu wichtigen Ressourcen zu sichern.
Hedonismus	Der Wert Hedonismus steht für Freude, Vergnügen und sinnliche Befriedigung für den Einzelnen. Er geht auf organismische Bedürfnisse zurück und die positive Erfahrung ihrer Erfüllung.
Stimulation	Erregung, Neuartigkeit und Herausforderungen im Leben bilden die motivationalen Grundlagen des Wertes Stimulation. Sein Ursprung liegt im Bedürfnis nach Abwechslung und optimaler Aktivierung.
Selbstbestimmung	Das motivationale Kernziel des Wertes Selbstbestimmung besteht in unabhängigem Denken und Handeln sowie im eigenständigen Wählen, Erschaffen und Erkunden. Dieser Wert leitet sich aus dem Streben nach Autonomie, Kontrolle und Meisterschaft ab.

8.3 Hierarchical Softmax

Originally developed by Morin and Bengio (2005) hierarchical softmax (HS) provides an efficient way for computing softmax values. The basic idea is to represent the vocabulary as the leafs of a binary tree and using the unique path from the root to each leaf to model the probability of a word. Prediction a word therefore amounts to predicting the sequence of decisions along its path. In other words instead of using matrix $W^{(2)}$ to calculate a probability for every word in V being the target word W_O we use the paths (which are represented by vectors) of the binary tree to calculate wether or not the leaf node is the target word. If the reached word is not the w_O we update the paths accordingly. Therefore in HS there is no output vector representation for every single word. Instead of applying the full output matrix $W^{(2)}$ to compute probabilities for all words in V, HS computes only the probabilities associated with the nodes on the path to the target word w_O. If the prediction does not reach w_O, the parameters along the traversed path are updated accordingly and a different path is taken the next time. Therefore, HS does not require explicit output vectors for every word saving enourmous computing effort.

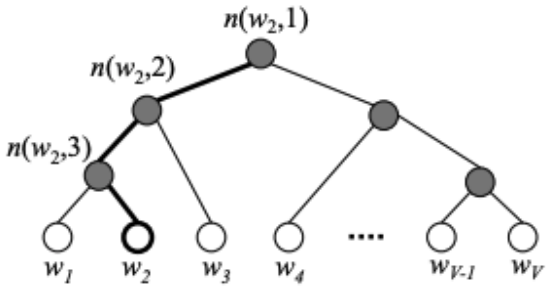

Figure 8.1: Figure copied from Rong (2014); An example binary tree used in the hierarchical softmax model. White nodes represent vocabulary words, while dark nodes denote internal decision nodes. The highlighted path illustrates the route from the root to the word w_2, where the path length is $L(w_2) = 4$. The notation $n(w, j)$ refers to the j-th node along the path from the root to the word w.

As mentioned each of the inner units has an output vector $\mathbf{v}'_{n(w,j)}$. With these vectors we can calculate the probability of a word being the target word w_O with:

$$p(w = w_O) = \prod_{j=1}^{L(w)-1} \sigma \quad n(w, j+1) = \text{ch}(n(w, j)) \quad \cdot \mathbf{v}'^{\top}_{n(w,j)} \mathbf{h}$$

Here, $\text{ch}(n)$ denotes the left child of node n. $\mathbf{v}'_{n(w,j)}$ is the vector representation (i.e., the output vector) of the internal node $n(w, j)$. $\mathbf{h}$ is the hidden-layer output(in the skip-gram model, $\mathbf{h} = \mathbf{v}_{w_I}$ and in the CBOW model, $\mathbf{h} = \frac{1}{C} \sum_{c=1}^{C} \mathbf{v}_{w_c}$).
The brackets $[\![\cdot]\!]$ denote a special indicator function, where $[\![x]\!]$ is defined

as

$$[\![x]\!] = \begin{cases} 1, & \text{if } x \text{ is true,} \\ -1, & \text{otherwise.} \end{cases}$$

The special function x checks whether the next node on the true path to the target word is the left child of the current inner node. At each step in the Huffman tree, it evaluates the binary expression

$$n(w, j+1) = \text{ch}(n(w, j)),$$

which is true if the correct path proceeds to the left child and false if it proceeds to the right child. If the equality is true, the function returns $+1$; if false, it returns -1. Multiplying the inner product by $+1$ or -1 flips the sign accordingly, ensuring that the sigmoid $\sigma(\cdot)$ assigns high probability to the correct branch and low probability to the incorrect one. This mechanism allows hierarchical softmax to encode each left/right decision while computing the probability of the target word.
The following example will make this calculation more understandable. Imagine we want to compute the probability for for w_2 being the output word w_O. The probability of w_2 is equal to the probability of a random walk starting from the root ending at the leaf unit of w_2. At each inner unit, we need to calculate the probabilities of going left or going right. The probability of going left at n is:

$$p(n, \text{left}) = \sigma \left({\mathbf{v}'_n}^\top \mathbf{h} \right)$$

Again $\mathbf{v}'_n$ is the vector representation of the inner unit and $\mathbf{h}$ is the vector representation of the input word determined by the hidden layer's. The probability of going right is simply 1 - the probability of going right:

$$p(n, \text{right}) = 1 - \sigma \left({\mathbf{v}'_n}^\top \mathbf{h} \right) = \sigma \left(-{\mathbf{v}'_n}^\top \mathbf{h} \right)$$

In Figure 1 you can see we that there are three inner units between the root and w_2 (the root included). Therefore we calculate the probability for w_2 simply by calculating three times whether or not we go left or right:

$$\begin{aligned} p(w_2 = w_O) &= p(n(w_2, 1), \text{left}) \cdot p(n(w_2, 2), \text{left}) \cdot p(n(w_2, 3), \text{right}) \\ &= \sigma \left({\mathbf{v}'_{n(w_2,1)}}^\top \mathbf{h} \right) \cdot \sigma \left({\mathbf{v}'_{n(w_2,2)}}^\top \mathbf{h} \right) \cdot \sigma \left(-{\mathbf{v}'_{n(w_2,3)}}^\top \mathbf{h} \right) \end{aligned}$$

which is just an applied formulation for the formula above:

$$p(w = w_O) = \prod_{j=1}^{L(w)-1} \sigma \left(n(w, j+1) = \text{ch}(n(w, j)) \right) {\mathbf{v}'_{n(w,j)}}^\top \mathbf{h}$$

Now we can derive the update equation for the vectors of the inner units. We again first start with one-word context model. For a single training instance the error function is

defined as: $n(w, j+1) = \text{ch}(n(w, j))$

$$E = -\log p(w = w_O \mid w_I) = -\sum_{j=1}^{L(w)-1} \log \sigma \quad (n(w, j+1) = \text{ch}(n(w, j)))\mathbf{v}_j'^{\top}\mathbf{h}$$

If we take the derivative of E with regard to $\mathbf{v}_j'\mathbf{h}$ we obtain:

$$\frac{E}{\mathbf{v}_j'^{\top}\mathbf{h}} = \quad \sigma((n(w, j+1) = \text{ch}(n(w, j)))\ \mathbf{v}_j'^{\top}\mathbf{h}) - 1 \quad (n(w, j+1) = \text{ch}(n(w, j)))$$

$$= \begin{array}{ll} \sigma\ \ \mathbf{v}_j'^{\top}\mathbf{h}\ \ - 1, & \text{if } (n(w, j+1) = \text{ch}(n(w, j))) = 1, \\ \sigma\ \ \mathbf{v}_j'^{\top}\mathbf{h}\ \ , & \text{if } (n(w, j+1) = \text{ch}(n(w, j))) = -1, \end{array}$$

$$= \sigma\ \ \mathbf{v}_j'^{\top}\mathbf{h}\ \ - t_j.$$

here t_j equals 1 if $n(w, j+1) = \text{ch}(n(w, j)) = 1$ and $t_j = 0$ otherwise.
Now we can take the derivative of E with regard to the vector representation of the inner units $n(w, j)$:

$$\frac{E}{\mathbf{v}_j'} = \frac{E}{(\mathbf{v}_j'^{\top}\mathbf{h})} \cdot \frac{(\mathbf{v}_j'^{\top}\mathbf{h})}{\mathbf{v}_j'} = \quad \sigma(\mathbf{v}_j'^{\top}\mathbf{h}) - t_j \ \ \mathbf{h}$$

from which we get the following update equation:

$$\mathbf{v}_j'^{(\text{new})} = \mathbf{v}_j'^{(\text{old})} - \quad \sigma(\mathbf{v}_j'^{\top}\mathbf{h}) - t_j \ \ \mathbf{h}$$

Which is applied to $j = 1, 2, \ldots, L(w) - 1$.We can interpret the term $\sigma(\mathbf{v}_j'^{\top}\mathbf{h}) - t_j$ as the prediction error associated with the inner node $n(w, j)$. Again, each inner node performs a binary decision: it must determine whether the correct path proceeds to the left or to the right child. A value of $t_j = 1$ indicates that the true branch is the left child, whereas $t_j = 0$ corresponds to the right child. The quantity $\sigma(\mathbf{v}_j'^{\top}\mathbf{h})$ represents the model's predicted probability of taking the left branch. When this prediction matches the ground truth, the resulting update to $\mathbf{v}_j'$ is small; when the prediction diverges, $\mathbf{v}_j'$ is adjusted-moved either closer to or farther from $\mathbf{h}$-to reduce the prediction error.
This update equation can be used for CBOW and Skip-Gram with slight modifications. When used for skip-gram we need to repeat this update procedure for each of the C words in the output context. In order to update the weight in $W^{(1)}$ we take the derivative of E with regards to the output of the hidden layer:

$$\frac{E}{\mathbf{h}} = \sum_{j=1}^{L(w)-1} \frac{E}{\mathbf{v}'_j\mathbf{h}} \cdot \frac{\mathbf{v}'_j\mathbf{h}}{\mathbf{h}}$$

$$= \sum_{j=1}^{L(w)-1} \sigma(\mathbf{v}'^{\top}_j \mathbf{h}) - t_j \quad \mathbf{v}'_j$$

$$:= EH$$

EH here again is the sum of all vectors each weighed by its prediction error. Therefore we can just pop it in the original update equation for the input -> hidden weights in the multi-word CBOW scenario:

$$\mathbf{v}^{(\text{new})}_{w_{I,c}} = \mathbf{v}^{(\text{old})}_{w_{I,c}} - \frac{1}{C} \quad EH^{\top}$$

In the case of skip-gram its similar. Here we calculate an EH value for each word in the skip-gram context and plug the sum of them into:

$$\mathbf{v}^{(\text{new})}_{w_I} = \mathbf{v}^{(\text{old})}_{w_I} - \quad EH^{\top}$$

Overall we can see that HS reduces the computational complexity per training instance per context word from $O(V)$ to $O(log(V))$ which is a huge gain in efficiency.

8.4 Negative Sampling

Another method to improve the computational efficiency of our training is Negative Sampling (NS). The core idea behind NS is quite simple: instead of updating all output vectors for every training instance, we update only a small subset of them. Specifically, we always include the positive sample-the true context word-and then select a few additional words as negative samples. The model should assign a high score to the positive sample and low scores to the negative ones. By focusing only on this small set, NS drastically reduces the number of required updates while still guiding the embeddings toward meaningful semantic structure. Following this logic the loss in NS can be calculated by the following formula:

$$E = -\log \sigma \quad \mathbf{v}'_{w_O}{}^{\top}\mathbf{h} \quad - \sum_{w_j \in W_{\text{neg}}} \log \sigma \quad -\mathbf{v}'_{w_j}{}^{\top}\mathbf{h}$$

w_O is as always the output word (the positive sample), $\mathbf{v}'_{wO}$ is the corresponding output vector; $\mathbf{h}$ is the output value of the hidden layer (in CBOW: $\mathbf{h} = \frac{1}{C} \sum_{c=1}^{C} \mathbf{v}_{w_c}$ and in skip-gram: $\mathbf{h}$=$\mathbf{v}_{wI}$). $\mathcal{W}_{\text{neg}} = \{w_j \mid j = 1, \ldots, K\}$ is the set of words comprising the negative sample.
In NS for us to get the update equations of the word vectors, we first take the derivative of the loss E with regards to the net input of the output unit w_j:

$$\frac{E}{{\mathbf{v}'_{w_j}}^\top \mathbf{h}} = \sigma \quad {\mathbf{v}'_{w_j}}^\top \mathbf{h} \quad - t_j.$$

where t_j is the "label" of the word w_j. t is either 1 when w_j is a positive sample or 0 if w_j is a negative sample. Now we can take the derivative of E with regard to the output vector of the word w_J,

$$\frac{E}{\mathbf{v}'_{w_j}} = \frac{E}{{\mathbf{v}'_{w_j}}^\top \mathbf{h}} \cdot \frac{{\mathbf{v}'_{w_j}}^\top \mathbf{h}}{\mathbf{v}'_{w_j}} = \quad \sigma({\mathbf{v}'_{w_j}}^\top \mathbf{h}) - t_j \quad \mathbf{h}$$

From this we can get the following update equation for its output vector:

$${\mathbf{v}'_{w_j}}^{(\text{new})} = {\mathbf{v}'_{w_j}}^{(\text{old})} - \quad \sigma({\mathbf{v}'_{w_j}}^\top \mathbf{h}) - t_j \quad \mathbf{h}$$

Now, the great improvement in computational efficiency in NS comes from the fact we only need to apply this update equation to $w_j \in \{ w_O \} \cup \mathcal{W}_{\text{neg}}$ instead of every word in the vocabulary. The update equation can be used the same way for CBOW and for Skip-Gram. For Skip-Gram we apply the equation for one context word at a time. Finally, to backpropagate the error of the hidden layer and updating the input vectors of the words we take the derivative of E with regard to the hidden layer s output:

$$\frac{E}{\mathbf{h}} = \sum_{w_j \in \{ w_O \} \cup W_{\text{neg}}} \frac{E}{{\mathbf{v}'_{w_j}}^\top \mathbf{h}} \cdot \frac{{\mathbf{v}'_{w_j}}^\top \mathbf{h}}{\mathbf{h}}$$

$$= \sum_{w_j \in \{ w_O \} \cup W_{\text{neg}}} \sigma({\mathbf{v}'_{w_j}}^\top \mathbf{h}) - t_j \quad \mathbf{v}'_{w_j} \quad := \quad EH$$

Now we again plug EH into the original update equation for the input -> hidden weights in the multi-word CBOW model:

$$\mathbf{v}_{w_{I,c}}^{(\text{new})} = \mathbf{v}_{w_{I,c}}^{(\text{old})} - \frac{1}{C} \quad EH^\top$$

and for Skip-Gram, where we again calculate EH for each word in the context, we plug the sum of the EH values into the original update equation from above:

$$\mathbf{v}_{w_I}^{(\text{new})} = \mathbf{v}_{w_I}^{(\text{old})} - \quad EH^\top$$

8.5 Anchored, weighted t-SNE neighborhood visualization

To visualize ideological semantic differences in a way that is comparable across models, we construct for each political term w a *single* two-dimensional neighborhood map

and then place the ideology-specific embedding of w into this shared 2D space. Because independently trained Word2Vec models are only comparable through an orthogonal transformation, we first align each ideology-specific embedding space to a common reference space via orthogonal Procrustes alignment (as introduced in the baseline methodology). Concretely, we first train an new Word2Vec model with the united data of our three ideological corpora. Then for each of our ideological model $m \in \{\text{left}, \text{center}, \text{right}\}$ we calculate the orthogonal rotation matrix R_m (with $R_m^\top R_m = I$) that maps ideological embeddings of w into the vector space of our united model. The aligned word vectors are then

$$\tilde{\mathbf{v}}_w^{(m)} = \mathbf{v}_w^{(m)} R_m \qquad \text{for } m \in \{\text{left}, \text{center}, \text{right}\},$$

while the word vector in the united model is $\mathbf{v}_w^{(\text{united})}$ directly.
We define the neighborhood vocabulary $N(w)$ by taking the top-n most similar words to w in the united model (here $n = 75$). For each neighbor $u \in N(w)$ we obtain its united embedding $\mathbf{v}_u$ and normalize all neighbor vectors to unit length,

$$\hat{\mathbf{v}}_u = \frac{\mathbf{v}_u}{\|\mathbf{v}_u\|_2},$$

so that dot products correspond to cosine similarity. We then compute a two-dimensional embedding for the neighbor set *once* using t-SNE with cosine distance, yielding coordinates $\mathbf{z}_u \in \mathbb{R}^2$ for all $u \in N(w)$. This produces a fixed 2D "neighbor cloud" that is shared across ideological variants of the target term.
To place each ideology-specific target occurrence into the fixed 2D neighbor geometry, we use a weighted anchored projection. For a given aligned word vector $\tilde{\mathbf{v}}_w$, we first normalize it,

$$\hat{\mathbf{v}}_w = \frac{\tilde{\mathbf{v}}_w}{\|\tilde{\mathbf{v}}_w\|_2},$$

and compute cosine similarities to all normalized neighbors:

$$s_u = \hat{\mathbf{v}}_u^\top \hat{\mathbf{v}}_w \qquad \text{for all } u \in N(w).$$

We then select the k' most similar neighbors (here $k' = 7$) as the local anchor set $A(w) \subseteq N(w)$. The anchored 2D position of the target is defined as a similarity-weighted centroid of the anchors' 2D coordinates. To emphasize the most similar anchors, similarities are converted to weights via a temperature-scaled softmax:

$$\alpha_u = \frac{\exp\left(s_u - \max_{a \in A(w)} s_a\right)/T}{\sum_{a \in A(w)} \exp\left(s_a - \max_{a' \in A(w)} s_{a'}\right)/T} \qquad \text{for } u \in A(w),$$

where T denotes the temperature parameter (here $T = 0.05$). Finally, the anchored target coordinate is

$$\mathbf{z}_w = \sum_{u \in A(w)} \alpha_u \, \mathbf{z}_u.$$

Applying this procedure to $\tilde{\mathbf{v}}_w^{(\text{left})}$, $\tilde{\mathbf{v}}_w^{(\text{center})}$, $\tilde{\mathbf{v}}_w^{(\text{right})}$, and $\mathbf{v}_w^{(\text{united})}$ yields four comparable target locations within the same fixed neighbor cloud. Differences between these anchored positions indicate ideological shifts in the local semantic neighborhood of w, while holding the 2D neighborhood geometry constant across groups.

Bibliography

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In Proceedings of the International Conference on Learning Representations.

Alina Arseniev-Koehler. 2024. Theoretical foundations and limits of word embeddings: What types of meaning can they capture? 53(4):1753–1793.

Alina Arseniev-Koehler and Jacob G Foster. 2022. Machine learning as a model for cultural learning: Teaching an algorithm what it means to be fat. Sociological Methods and Research, 51(4):1484–1539.

Deepak Suresh Asudan, Naresh Kumar Nagwan, and Pradeep Singh. 2023. Impact of word embedding models on text analytics in deep learning – a review. Machine Learning with Applications, 12:100523.

Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh. 2023. Impact of word embedding models on text analytics in deep learning environment: a review. Artificial intelligence review, 56(9):10345–10425.

Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. 2017. Words are malleable: Computing semantic shifts in political and media discourse. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1509–1518.

Marina Barnea. 2003. Personal values and party orientations in different cultures. Ph.D. thesis, The Hebrew University of Jerusalem, Israel.

Marina F Barnea and Shalom H Schwartz. 1998. Values and voting. Political Psychology, 19(1):17–40.

Nicolas Frenzel Baudisch. 2018. Individuen mit widersprüchlichen Wertvorstellungen. Springer.

Andreas Blank and Peter Koch. 1999. Historical Semantics and Cognition. Walter de Gruyter.

Leonard Bloomfield. 1933. Language. Allen & Unwin.

Diana Boer and Ronald Fischer. 2013. How and when do personal values guide our attitudes and sociality? Explaining cross-cultural variability in attitude–value linkages. Psychological Bulletin, 139(5):1113.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4758–4781, Online. Association for Computational Linguistics.

Andrei Boutyline and Alina Arseniev-Koehler. 2025. Meaning in hyperspace: Word embeddings as tools for cultural measurement. Annual Review of Sociology, 51.

Andrei Boutyline and Ethan Johnston. 2023. Forging better axes: Evaluating and improving the measurement of semantic dimensions in word embeddings. OSF preprint.

Daniel Braun. 2022. Tracking semantic shifts in German court decisions with diachronic word embeddings. In Proceedings of the Natural Legal Language Processing Workshop 2022, pages 218–227.

Rogers Brubaker. 2020. Populism and Nationalism. Nations and Nationalism, 26(1):44–66.

Michel Bréal. 1899. Essai de sémantique. Hachette, Paris.

Gian Vittorio Caprara, Shalom Schwartz, Cristina Capanna, Michele Vecchione, and Claudio Barbaranelli. 2006. Personality and politics: Values, traits, and political choice. Political Psychology, 27(1):1–28.

Thomas M Carsey and Geoffrey C Layman. 2006. Changing sides or changing minds? Party identification and policy preferences in the american electorate. American Journal of Political Science, 50(2):464–477.

Ilias Chalkidis and Stephanie Brandl. Llama meets EU: Investigating the european political spectrum through the lens of LLMs. Preprint, arxiv:2403.13592 [cs].

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In Proceedings of the International AAAI conference on web and social media, volume 11, pages 512–515.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. Journal of the American society for Information science, 41(6):391–407.

Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. Following the autoregressive nature of LLM embeddings via compression and alignment. Preprint, arxiv:2502.11401 [cs].

Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Following the autoregressive nature of llm embeddings via compression and alignment. arXiv preprint arXiv:2502.11401.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, volume 1 (long and short papers), pages 4171–4186.

David Dubin. 2004. The most influential paper Gerard Salton never wrote. Graduate School of Library and Information Science, University of Illinois.

Haim Dubossarsky, Yulia Tsvetkov, Chris Dyer, and Eitan Grossman. 2015. A bottom up approach to category mapping and meaning change. In NetWordS, pages 66–70.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1136–1145.

Robert M. Entman. 1993. Framing: Towards clarification of a fractured paradigm. In McQuail's Reader in Mass Communication Theory, pages 390–397.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. arXiv preprint arXiv:1411.4166.

Jaouhar Fattahi, Feriel Sghaier, Mohamed Mejri, Ridha Ghayoula, Sahbi Bahroun, and Marwa Ziadia. 2024. Sexism discovery using CNN, word embeddings, NLP and data augmentation. In 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), pages 1685–1690. IEEE.

John R. Firth. 1957. Papers in Linguistics. Oxford University Press, London.

Michael Freeden. 1996. Ideologies and political theory: A conceptual approach. Clarendon Press.

Matthew Freestone and Shubhra Kanti Karmaker Santu. 2024a. Word embeddings revisited: Do llms offer something new. arXiv preprint arXiv:2402.11094.

Matthew Freestone and Shubhra Kanti Karmaker Santu. 2024b. Word embeddings revisited: Do llms offer something new? arXiv preprint arXiv:2402.11094.

Walter Bryce Gallie. 1955. Essentially contested concepts. In Proceedings of the Aristotelian society, volume 56, pages 167–198.

Peter Gärdenfors. 2004. Conceptual spaces: The geometry of thought. MIT Press.

Peter Gärdenfors. 2011. Semantics based on conceptual spaces. In Proceedings of the Indian Conference on Logic and Its Applications, pages 1–11. Springer.

Bernhard Giesen and Robert Seyfert. 1999. Kollektive Identität. Suhrkamp Frankfurt a. M.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. arXiv preprint arXiv:2004.14118.

Joachim Grzega and Marion Schoener. 2007. English and General Historical Lexicology. Katholische Universität Eichstätt-Ingolstadt, Eichstätt-Ingolstadt.

Yanzhu Guo, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. How COVID-19 is changing our language : Detecting semantic shift in Twitter Word Embeddings. Preprint, arXiv:2102.07836.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. Social Justice Research, 20(1):98–116.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural shift or linguistic drift? Comparing two computational measures of semantic change. In Proceedings of the conference on empirical methods in Natural Language processing., volume 2016, page 2116.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Diachronic word embeddings reveal statistical laws of semantic change. ArXiv preprint arXiv:1605.09096.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JeSemE: Interleaving semantics and emotions in a web service for the exploration of language change phenomena. In Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 10–14, Santa Fe, New Mexico. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4479–4488. Association for Computational Linguistics.

Ronald Inglehart. 2015. The silent revolution: Changing values and political styles among Western publics. Princeton University Press.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political ideology detection using recursive neural networks. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (volume 1: long papers), pages 1113–1122.

S. Joshua Johnson, M. Ramakrishna Murty, and I. Navakanth. A detailed review on word embedding techniques with emphasis on word2vec. 83(13):37979–38007.

S Joshua Johnson, M Ramakrishna Murty, and I Navakanth. 2024. A detailed review on word embedding techniques with emphasis on word2vec. Multimedia Tools and Applications, 83(13):37979–38007.

Jason Jones, Ruhul Amin, Jessica Kim Mohammad, and Steven Skiena. 2020. Stereotypical gender associations in language have decreased over time. Sociological Science, 7:1–35.

John T Jost, Brian A Nosek, and Samuel D Gosling. 2008. Ideology: Its resurgence in social, personality, and political psychology. Perspectives on Psychological Science, 3(2):126–136.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. How do moral emotions shape political participation? a cross-cultural analysis of online petitions using language models. In Findings of the Association for Computational Linguistics ACL 2024, pages 16274–16289. Association for Computational Linguistics.

Kathleen Knight. 2006. Transformations of the concept of ideology in the twentieth century. American Political Science Review, 100(4):619–626.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. American Sociological Review.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In Proceedings of the 24th international conference on world wide web, pages 625–635.

Andrey Kutuzov. 2020. Distributional Word Embeddings in Modeling Diachronic Semantic Change. Ph.D. thesis, University of Oslo.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

George Lakoff. 2022. Moral politics: How liberals and conservatives think. University of Chicago Press.

Thomas K Landauer and Susan T Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review, 104(2):211.

Rémi Philippe Lebret. 2016. Word embeddings for natural language processing. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne.

Adrienne Lehrer. 1985. The influence of semantic fields on semantic change. Historical semantics, historical word formation, 29:283.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Saranya M and Amutha B. A survey of machine learning technique for topic modeling and word embedding. In 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), pages 1–6. IEEE.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Yash Mahajan, Matthew Freestone, Sathyanarayanan Aakur, and Shubhra Kanti Karmaker. 2025. Revisiting word embeddings in the llm era. arXiv preprint arXiv:2502.19607.

Yulia Maslennikova and Vladimir Bochkarev. Evaluation of word embedding models used for diachronic semantic change analysis. 2701(1):012082.

Yulia Maslennikova and Vladimir Bochkarev. 2024. Evaluation of word embedding models used for diachronic semantic change analysis. In Journal of Physics: Conference Series, volume 2701, page 012082. IOP Publishing.

Scott McDonald and Michael Ramscar. 2001. Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In Proceedings of the 23rd Annual Meeting of the Cognitive Science Society, pages 693–698, Edinburgh, Scotland. Cognitive Science Society.

Michael Calvin McGee. 1980. The "ideograph": A link between rhetoric and ideology. Quarterly journal of speech, 66(1):1–16.

Rada Mihalcea and Vivi Nastase. 2012. Word epoch disambiguation: Finding how words change over time. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 259–263, Jeju Island, Korea. Association for Computational Linguistics.

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

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. ACM Transactions on Internet Technology (TOIT), 17(3):1–23.

John W. Mohr. Measuring meaning structures. 24(1):345–370.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In International workshop on artificial intelligence and statistics, pages 246–252. PMLR.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 142–148, San Diego, California. Association for Computational Linguistics.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers, 36(3):402–407.

Allen Newell. 1980. Physical symbol systems. Cognitive science, 4(2):135–183.

Jonas Nölle, Stefan Hartmann, and Peeter Tinits. 2020. Language evolution research in the year 2020: A survey of new directions. Language Dynamics and Change, 10(1):3–26.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35:27730–27744.

Alexandr Pak, Atabay Ziyaden, Timur Saparov, Iskander Akhmetov, and Alexander Gelbukh. 2024. Word embeddings: A comprehensive survey. Research in Computing Science, 179(4):2005–2030.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Ludovic Rheault, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. Measuring emotion in parliamentary debates with automated textual analysis. 11(12):e0168843.

Milton Rokeach. 1973. The nature of human values. Free press.

Xin Rong. 2014. Word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.

Magnus Sahlgren. 2008. The distributional hypothesis. Italian Journal of linguistics, 20:33–53.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5):513–523.

Gerard Salton, Anita Wong, and Chung-Shu Yang. 1975. A vector space model for automatic indexing. Communications of the ACM, 18(11):613–620.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, and 1 others. 2021. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207.

Edward Sapir. 1921. An introduction to the study of speech. Language, 1(1):15.

Ferdinand de Saussure. 1983. Course in General Linguistics. Duckworth, London.

Hinrich Schütze. 1992. Word space. Advances in neural information processing systems, 5.

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

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Advances in Experimental Social Psychology, volume 25, pages 1–65. Elsevier.

Shalom H Schwartz. 2012a. Basic personal values and political orientations. Improving public opinion surveys: Interdisciplinary innovation and the American national election studies, 53:63–82.

Shalom H Schwartz. 2012b. An overview of the schwartz theory of basic values. Online readings in Psychology and Culture, 2(1):11.

Murray Shanahan. 2024. Talking about large language models. Communications of the ACM, 67(2):68–79.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022a. Political ideology and polarization: A multi-dimensional approach. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 231–243.

Barea Sinno, Bernardo Oviedo, Katherine Atwell, Malihe Alikhani, and Junyi Jessy Li. 2022b. Political ideology and polarization: A multi-dimensional approach. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 231–243. Association for Computational Linguistics.

Maximilian Spliethöver, Maximilian Keiff, and Henning Wachsmuth. 2022. No word embedding model is perfect: Evaluating the representation accuracy for social bias in the media. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 2081–2093. Association for Computational Linguistics.

Lukas Stankevicius and Mantas Lukoševicius. 2024. Extracting sentence embeddings from pretrained transformer models. Applied Sciences, 14(19):8887.

Gustaf Stern. 1931. Meaning and Change of Meaning: With Special Reference to the English Language. Wettergren & Kerbers.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen. 2021. Computational approaches to semantic change. Zenodo.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In Proceedings of the international AAAI conference on web and social media, volume 4, pages 178–185.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Dinithi Vithanage, Ping Yu, Lei Wang, and Chao Deng. Contextual word embedding for biomedical knowledge extraction: a rapid review and case study. 8(1):158–179.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. 2022. What language model architecture and pretraining objective works best for zero-shot generalization? In International Conference on Machine Learning, pages 22964–22984. PMLR.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alexander Smola, and Josh Attenberg. 2009. Feature hashing for large scale multitask learning. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1113–1120. ACM.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In Proceedings

of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 23–30, Online. Association for Computational Linguistics.

Naman Bansal Yash Mahajan, Matthew Freestone. 2025. Revisiting word embeddings in the llm era. arXiv preprint arXiv:2402.11094.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. ArXiv, abs/2303.18223.