Improving Interpretability of Lexical Semantic Change with Neurobiological Features

Kohei Oda, Hiroya Takamura, Kiyoaki Shirai, Natthawut Kertkeidkachorn

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Improving Interpretability of Lexical Semantic Change with Neurobiological Features

Kohei Oda[1] **Hiroya Takamura**[2] **Kiyoaki Shirai**[1] **Natthawut Kertkeidkachorn**[1]
[1]Japan Advanced Institute of Science and Technology
[2]National Institute of Advanced Industrial Science and Technology
[1]{s2420017,kshirai,natt}@jaist.ac.jp
[2]takamura.hiroya@aist.go.jp

Abstract

Lexical Semantic Change (LSC) is the phenomenon in which the meaning of a word change over time. Most studies on LSC focus on improving the performance of estimating the degree of LSC, however, it is often difficult to interpret how the meaning of a word change. Enhancing the interpretability of LSC is a significant challenge as it could lead to novel insights in this field. To tackle this challenge, we propose a method to map the semantic space of contextualized embeddings of words obtained by a pre-trained language model to a neurobiological feature space. In the neurobiological feature space, each dimension corresponds to a primitive feature of words, and its value represents the intensity of that feature. This enables humans to interpret LSC systematically. When employed for the estimation of the degree of LSC, our method demonstrates superior performance in comparison to the majority of the previous methods. In addition, given the high interpretability of the proposed method, several analyses on LSC are carried out. The results demonstrate that our method not only discovers interesting types of LSC that have been overlooked in previous studies but also effectively searches for words with specific types of LSC.[1]

1 Introduction

The meanings of words change over time. For example, according to the Oxford English Dictionary (OED)[2], the word *gay* acquired the meaning of *homosexual* around 1934, in addition to its earlier meaning of *cheerful*. This phenomenon is called Lexical Semantic Change (LSC) and actively studied in recent years (Tahmasebi et al., 2019, 2021; Periti and Montanelli, 2024). Many studies in this field represent the meanings of words as vectors using embedding models, such as static word embeddings (Mikolov et al., 2013) and BERT (Devlin et al., 2019), and learn separate spaces for different time periods (Kim et al., 2014; Hamilton et al., 2016; Bamler and Mandt, 2017) or handle multiple time periods within the same space (Hu et al., 2019; Giulianelli et al., 2020; Martinc et al., 2020b). While these techniques are useful for estimating the degree of LSC, they are inappropriate for humans to interpret LSC.

Several methods have been proposed to improve the interpretability of LSC, including a method presenting neighboring words in a vector space (Gonen et al., 2020), obtaining representative co-occurrence words (Montariol et al., 2021), predicting substitutions (Card, 2023), assigning predefined word senses (Tang et al., 2023), and generating definition sentences of word meanings (Giulianelli et al., 2023; Fedorova et al., 2024). These methods help humans interpret LSC through natural language, e.g., by showing indicative words or definition sentences. However, explanations of LSC based on words and sentences are ambiguous and lack a systematic explanatory framework.

Motivated by the above, we propose a method to improve the interpretability of LSC by using neurobiological features proposed by Binder et al. (2016), which we call *Binder features* in this paper. There are 65 Binder features such as *Vision*, *Audition*, and *Happy*. The values of these 65 Binder features have been estimated for 535 English words and are open to the public.[3] Based on previous studies (Utsumi, 2018, 2020; Turton et al., 2021), we use the public dataset above to train a regression model that maps the BERT semantic space to the Binder space for the quantitative and multi-perspective interpretation of LSC.

First, the potential of the Binder features in an-

[1]Our code is available at: https://github.com/iehok/LSC_with_Binder.
[2]https://www.oed.com/
[3]https://www.neuro.mcw.edu/index.php/resources/brain-based-semantic-representations/

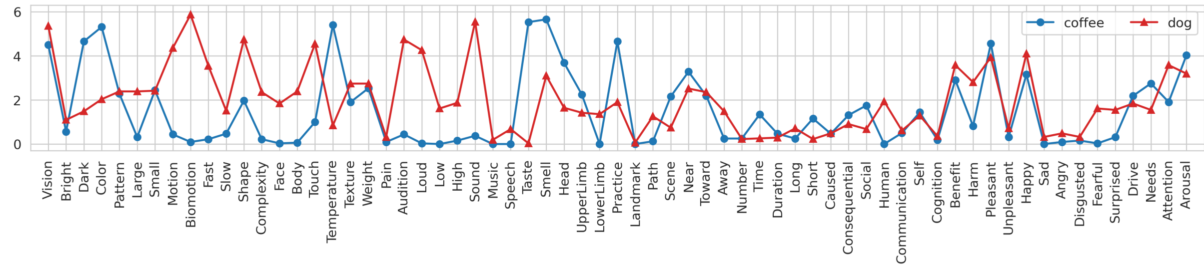


Figure 1: Binder feature values for "coffee" and "dog"

alyzing LSC is evaluated by applying our method to a task aimed at estimating the degree of LSC. Second, utilizing the high interpretability of our method, we analyze types of LSC. The integration of our method with Sparse PCA (Principal Component Analysis) enables us to identify interesting types of LSC that have not been found in previous studies. Finally, we apply our method to detect amelioration and pejoration (Traugott, 2017), and successfully identify ameliorative and pejorative words in a real corpus.

The contributions of our paper are summarized as follows:

- We introduce neurobiological features into the field of lexical semantic change, thereby improving the interpretability.
- We discover several interesting types of LSC that have not been noted in previous studies by combining our method with Sparse PCA.
- We propose a method that can easily detect specific types of LSC, amelioration and pejoration, using our approach.

2 Related Work

2.1 Lexical Semantic Change

LSC is mainly studied in the fields of linguistics and natural language processing (NLP). Even when being constrained to NLP, numerous methods are proposed such as a method that utilizes mutual information (Gulordava and Baroni, 2011; Hamilton et al., 2016; Schlechtweg et al., 2019), Bayesian models (Emms and Jayapal, 2016; Frermann and Lapata, 2016; Inoue et al., 2022), and static word embeddings (Kulkarni et al., 2015; Takamura et al., 2017; Del Tredici et al., 2019).

Recently, with the emergence of pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) that can generate representations of the meanings of words in a context, methods using these models have been actively studied (Kutuzov and Giulianelli, 2020; Martinc et al., 2020a; Liu et al., 2021b). Hu et al. (2019) propose a method to identify how the meaning of a word changes by calculating the distribution of word senses over time, where example sentences in the OED are used to assign senses to words in the corpus. Giulianelli et al. (2020) propose a method to calculate the distribution of usage types (pseudo senses) without using a dictionary. This method uses k-means clustering on a set of contextualized embeddings from all time periods to assign usage types to words in the corpus. Additionally, the degree of LSC between two different time periods is estimated using either the Jensen-Shannon divergence (JSD) between usage type distributions or the average pairwise distance (APD) between sets of contextualized embeddings from these time periods. In this study, we model LSC based on Giulianelli et al. (2020) coupled with the Binder features to improve the interpretability.

2.2 Interpretable Word Embeddings

Interpretable representations, i.e., methods of assigning roles (interpretations) to each dimension of an embedding, have been extensively studied (Panigrahi et al., 2019; Şenel et al., 2020; Aloui et al., 2020). However, these methods often suffer from a lack of clarity of a role for each dimension or coarse granularity of roles.

Binder et al. (2016) propose interpretable word vectors by defining 65 features based on neurobiological perspectives and manually assign these strengths (0 to 6) to 535 words, including 434 nouns, 62 verbs, and 39 adjectives. Figure 1 shows the Binder features and their corresponding values for "coffee" and "dog." The values of *Taste* and *Smell* are relatively high for "coffee," while the values of *Biomotion* and *Sound* are high for "dog." Additionally, the *Vision* feature has high values for both words.

Binder features are actively studied in the fields of cognitive linguistics and NLP. Utsumi (2020); Chersoni et al. (2021); Flechas Manrique et al. (2023) investigate what kind of information is encoded in static word embeddings, such as SGNS (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), by mapping these word embedding spaces to the Binder feature space. Turton et al. (2020) assign the Binder values to words other than the original 535 words by the aforementioned mapping from word embeddings. Turton et al. (2021) demonstrate that contextualized word embeddings generated from Transformer (Vaswani et al., 2017) based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) can derive the Binder values in the same way as the mapping of static word embeddings.

3 Mapping BERT Space to Binder Space

To enhance the interpretability of LSC, we first convert the semantic space of contextualized word embeddings derived from BERT to that of the Binder features. Specifically, a regression model is trained, which maps the BERT space (768 dimensions) to the Binder space (65 dimensions). The regression model, designated as ψ, is formalized as follows,

$$\mathbf{b}_w = \psi(\mathbf{r}_w), \tag{1}$$

where $\mathbf{r}_w$ and $\mathbf{b}_w$ are BERT and Binder vectors, respectively.

3.1 Word Embeddings on BERT Space

Let $\mathcal{C}$ be the corpus used for training, and let $\mathcal{C}_w$ be the set of (s, i), a pair of a sentence s in $\mathcal{C}$ that contains the target word w and its position i in s. The representation of w in the entire $\mathcal{C}$ is defined as follows.

$$\mathbf{r}_w = \frac{1}{|\mathcal{C}_w|} \sum_{(s,i)\in\mathcal{C}_w} \phi(s, i). \tag{2}$$

The function $\phi(s, i)$ denotes the hidden state of the final layer for the i-th token of the BERT model when s is given as an input. In this study, `bert-base-uncased`[4] is used as the BERT model.

The Clean Corpus of Historical American English (CCOHA) (Alatrash et al., 2020) is used as $\mathcal{C}$. CCOHA is an English corpus covering the period from 1820 to 2020, divided into ten-year segments. It consists of five genres: TV/Movies, Fiction, Magazine, Newspaper, and Non-fiction.

[4] https://huggingface.co/google-bert/bert-base-uncased

	LT	MLP
1910-2010	.571	.645
1960-2010	**.569**	.689

Table 1: Average MSE for 10 trials

3.2 Training of Regression Model

Two architectures of the regression model are applied: a simple linear transformation (LT) and a multilayer perceptron (MLP). The MLP consists of four hidden layers (300, 200, 100, 50 dimensions), following Turton et al. (2021). The output of each layer is activated by ReLU. To match the scale of the Binder value, in both the LT and the MLP, the final output is activated by Sigmoid and subsequently multiplied by 6 to convert values within the range of 0 to 6. The regression models are trained using 535 words associated with the Binder feature vectors (Binder et al., 2016). The loss function is set to the mean squared error (MSE) between predicted and ground-truth values of all Binder features of the target words.

3.3 Settings

We conduct experiments using two different periods of CCOHA: 1910-2010 and 1960-2010. The period 1910-2010 follows the setting in Giulianelli et al. (2020), while the period 1960-2010 is set to the most recent half of it, as the Binder dataset (Binder et al., 2016) was created in 2016. The performance of the trained regression model is evaluated using k-fold cross-validation, where k is set to 10. The batch size, the learning rate, and the number of epochs are set to 16, 1e-3 and 100, respectively. The quality of the regression model is evaluated by the MSE on the test set. The MSE is measured at each epoch, and the minimum MSE is recorded.

3.4 Results

Table 1 shows the average MSE for 10 trials of the cross-validation. LT significantly outperforms MLP, while the time period of the training corpus has a minimal influence on the results. This may be because the words in the Binder dataset are well-known and common, which leads to a relatively stable representation over time. This finding partially agrees with the results obtained by Hamilton et al. (2016).

4 Lexical Semantic Change Detection

This section proposes and evaluates a method to detect LSC using the Binder feature vectors.

4.1 Task Definition

SemEval-2020 Task 1 (Subtask 2) (Schlechtweg et al., 2020) is a task that aims to predict the degree of LSC of a word w. Specifically, the goal is to predict an LSC score representing how drastically the meaning of w changes between $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$, which are corpora of two different periods t_1 and t_2. The dataset consists of 37 English target words with manually assigned LSC scores. $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$ are parts of CCOHA from 1810 to 1860 and 1960 to 2010, respectively. Evaluation is performed by measuring Spearman's rank correlation coefficient between the predicted and ground-truth LSC scores.

4.2 Predicting Degree of LSC

To predict the degree of LSC of a word w from t_1 to t_2, the set of contextualized embeddings of w in the corpus $\mathcal{C}^t_w$ is calculated for each time period:

$$\mathcal{U}^t_w = \bigcup_{(s,i)\in\mathcal{C}^t_w} \{\,\psi(\phi(s,i))\,\}, \tag{3}$$

where ψ is the regression model, either LT or MLP explained in Section 3. Then, following Giulianelli et al. (2020), the degree of LSC between $\mathcal{U}^{t_1}_w$ and $\mathcal{U}^{t_2}_w$ is measured by the average pairwise distance (APD):

$$\mathrm{APD}(\mathcal{U}^{t_1}_w, \mathcal{U}^{t_2}_w) = \frac{1}{|\mathcal{U}^{t_1}_w| \cdot |\mathcal{U}^{t_2}_w|} \sum_{\mathbf{u}_i\in\mathcal{U}^{t_1}_w} \sum_{\mathbf{u}_j\in\mathcal{U}^{t_2}_w} d(\mathbf{u}_i, \mathbf{u}_j), \tag{4}$$

where d is a distance function. We compared the following three distance functions: Euclidean distance, cosine distance, and Spearman distance. Spearman distance is defined as $(1 - \mathrm{sc}(\mathbf{u}_i, \mathbf{u}_j))$, where $\mathrm{sc}(\mathbf{u}_i, \mathbf{u}_j)$ is Spearman's rank correlation coefficient between sets of values of dimensions in two vectors.

4.3 Results

Table 2 shows a comparison between the baseline BERT space and our methods based on different regression models. The performance of LSC detection is slightly improved by mapping to the Binder space using the linear regression model. The architecture of the regression model significantly impacts the performance of LSC detection, while the period of the corpus used for training the regression model has a relatively small impact; this tendency is similar to that in Table 1. Among the three distance functions, the cosine distance is relatively stable and performs well.

Model	Euclid	Cosine	Spearman
BERT space	.616	.645	.618
LT-1910-2010	.633	.644	**.647**
LT-1960-2010	**.635**	**.667**	.634
MLP-1910-2010	.499	.587	.562
MLP-1960-2010	.483	.442	.540

Table 2: Spearman's rank correlation coefficient for SemEval-2020 Task 1. "BERT space" means a method that calculates APD in the BERT space without mapping it to the Binder space.

Model	EK	Score
SSCS (Tang et al., 2023)	✓	.589
XL-LEXEME (Cassotti et al., 2023)	✓	.757
SDML (Aida and Bollegala, 2024)	✓	.774
NLPCR (Rother et al., 2020)		.436
APD (Laicher et al., 2021)		.571
ScaledJSD (Card, 2023)		.547
SSCD (Aida and Bollegala, 2023)		.383
LT-1960-2010 (ours)		**.667**

Table 3: Spearman's rank correlation coefficient for previous methods on SemEval-2020 Task 1. "EK" (external knowledge) means methods that are fine-tuned with WiC corpora (Raganato et al., 2020; Martelli et al., 2021; Liu et al., 2021a) or methods using the information of dictionaries such as WordNet (Miller, 1994) and BabelNet (Navigli and Ponzetto, 2010).

Table 3 shows a comparison of our method with other existing methods. Although our method is simple, it achieves the best performance compared to other methods that do not use any external knowledge.

5 Analysis of LSC Types

This section describes an analysis of LSC types using our method. Given the high interpretability of neurobiological features, our goal is to identify the types of semantic changes of words between two different periods t_1 and t_2.

5.1 Target Words

The target words used for this analysis are collected from WordNet (Miller, 1994) lemmas meeting the following three conditions: (i) included in the vocabulary of the BERT tokenizer, (ii) having two

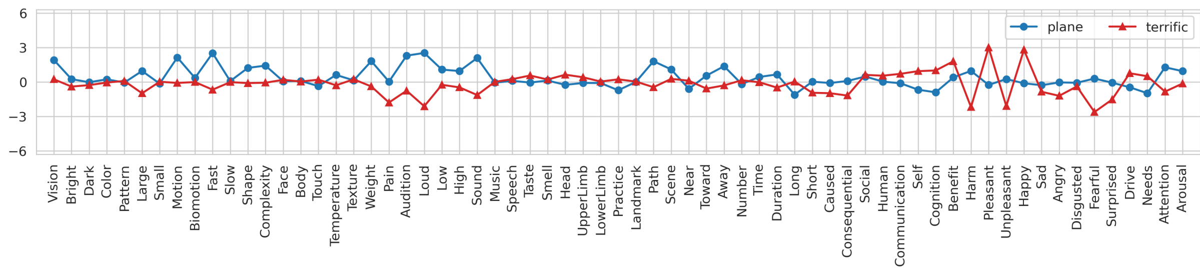


Figure 2: LSC vectors for *plane* and *terrific*

or more senses in WordNet, and (iii) four or more characters long. Condition (i) is set because our method is incapable of handling words that are divided into subwords. Condition (ii) is set because words whose meanings have changed are likely to have newly added senses, thereby resulting in polysemy. Condition (iii) is set because short words are more likely to become subwords within other words. Based on these three conditions, a total of 8,570 target words are chosen.

5.2 Corpora

The CCOHA 1910s (from 1910 to 1920) and 2000s (from 2000 to 2010) are used as the corpora $\mathcal{C}^{t_1}$ and $\mathcal{C}^{t_2}$. This corresponds to the first and last decades of the period from 1910 to 2010 used for training the regression model (Section 3).

5.3 Methods

To analyze how the meaning of words changed between two periods t_1 and t_2, the LSC vector of the word w, denoted as $\mathbf{v}_{\text{lsc}}(w)$, is computed as follows:

$$\mathbf{v}_{\text{lsc}}(w) = \frac{1}{|\mathcal{U}_w^{t_2}|} \sum_{\mathbf{u}_i \in \mathcal{U}_w^{t_2}} \mathbf{u}_i - \frac{1}{|\mathcal{U}_w^{t_1}|} \sum_{\mathbf{u}_i \in \mathcal{U}_w^{t_1}} \mathbf{u}_i. \quad (5)$$

This vector represents the semantic changes of all Binder features. A positive value in a dimension of the LSC vector means that the meaning of the corresponding Binder feature is newly acquired from t_1 to t_2, while a negative value implies a loss of the meaning.

After calculating the LSC vectors for all target words, Sparse PCA is applied to the LSC vectors of the 500 target words with the largest norms, supposing that the meanings of words with small norms are not significantly changed. Unlike conventional PCA, Sparse PCA enhances interpretability by setting many elements in the eigenvectors to zero, and the eigenvectors do not need to be orthogonal to each other. Since the number of principal components should be predetermined, it is set to 10 in this experiment. The analysis of different numbers of principal components remains a subject for future work.

It is hypothesized that each principal component (PC) of Sparse PCA represents a type of LSC. For each PC, the top three Binder features with the highest values in the eigenvector are extracted to provide a clear interpretation of the LSC type. Subsequently, we check the words in descending or ascending order of their values of the principal component and verify whether they are representative words. The validity of the chosen representative words is evaluated by the following procedures. First, following Giulianelli et al. (2020), usage types (pseudo senses) are assigned to the target words in example sentences by conducting k-means clustering on a set of contextualized embeddings. Second, the five examples closest to the center of each cluster are examined to confirm whether they are correctly divided according to their meanings. Finally, the change in the distribution of usage types from t_1 to t_2 is checked to investigate whether it supports the LSC type augmented by the related Binder features.

This method is similar to the analysis by applying PCA in BERT space (Aida and Bollegala, 2025), but enhances the interpretability of LSC types. First, not only words with large or small principal components but also the values of eigenvectors can be used for analyzing LSC types. Second, since Sparse PCA assigns a zero to many elements, it is easier to find relevant (non-zero) Binder features for each LSC type.

5.4 Results

Figure 2 shows the LSC vectors for *plane* and *terrific*. According to the OED, the word *plane* acquired the meaning of *airplane* around 1908, in addition to its existing meaning of *a flat geometri-*

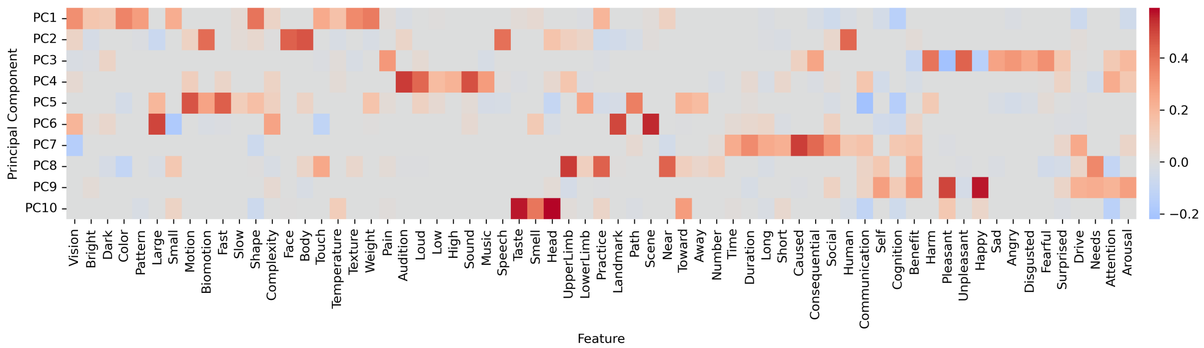


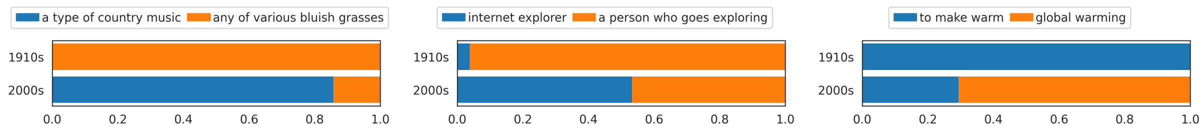
Figure 3: Eigenvectors obtained by Sparse PCA. The number of principal components is set to 10. The horizontal and vertical axes represent the 65 Binder features and the 10 principal components, respectively.

PC	LSC Type Label	Top 3 Binder Features	Representative Words
1	Artifact	Shape, Weight, Color	↑ console, plastic, vogue ↓ overall, album, bluegrass
2	Human	Body, Face, Human	↑ coach, shooter, racer ↓ yahoo, explorer, major
3	Negative Meaning	Unpleasant, Harm, Fearful	↑ serial, aids, parkinson ↓ offence, terrific, crook
4	Sound	Audition, Sound, Loud	↑ bluegrass, plane, blues ↓ instrumentation, click, booming
5	Transportation	Motion, Fast, Path	↑ pickup, sedan, plane ↓ steamed, omnibus, coach
6	Place	Scene, Large, Landmark	↑ facility, resort, berkeley ↓ manila, chihuahua, bologna
7	Social	Caused, Consequential, Duration	↑ warming, launch, summit ↓ briefs, console, offensive
8	Familiar Thing	UpperLimb, Practice, Near	↑ topical, sink, blackberry ↓ album, shooter, warming
9	Positive Meaning	Happy, Pleasant, Benefit	↑ bonding, outgoing, terrific ↓ intelligence, utility, console
10	Food	Head, Taste, Smell	↑ bologna, bourbon, steamed ↓ alcoholic, blackberry, bluegrass

Table 4: The result of analysis of Sparse PCA. "LSC Type Label" is a manually assigned label for the LSC type. The symbols ↑ and ↓ indicate words with relatively large and small principal component values, respectively, suggesting the words have acquired or lost their meanings of the features.

cal surface. As illustrated in Figure 2, the values of the Binder features such as *Motion*, *Audition*, and *Path* exhibit a substantial increase. Additionally, according to the OED, *terrific* acquired the meaning of *amazing* around 1871, in addition to its existing meaning of *causing terror*. The values of the Binder features *Pleasant* and *Happy* have increased significantly, while the values of the Binder features *Harm*, *Unpleasant*, and *Fearful* have decreased significantly. This indicates that the major meaning of *terrific* has shifted from a negative to a positive meaning.

Figure 3 shows the eigenvectors obtained by Sparse PCA. Many elements in the eigenvectors are zero, making them relatively easy to interpret. In addition, by examining the absolute values in the eigenvectors, it is possible to identify Binder features that are deeply related to or not related to LSC. For example, the absolute values of *Vision* at the first, sixth, and seventh PCs are relatively high, indicating *Vision* is likely to be deeply related to LSC. On the other hand, the absolute values of *Number* are nearly zero in all PCs, suggesting that *Number* does not contribute to LSC.



(a) bluegrass (b) explorer (c) warming

Figure 4: Distributions of the usage types for *bluegrass*, *explorer*, and *warming*

Table 4 shows the types of LSC corresponding to PCs. Many interesting types of LSC, which have not been noted in previous studies (Traugott, 2017; Campbell, 2020), are discovered by our method. Figure 4 shows the distributions of the usage types for some illustrative examples of words: *bluegrass*, *explorer*, and *warming*. For *bluegrass* in PC4, the meaning has changed from *any of various bluish grasses* to *a type of country music*, shifting to a meaning related to sounds. For *explorer* in PC2, the meaning related to humans has declined due to its increased use in the collocation *internet explorer*. For *warming* in PC7, the meaning has changed to a meaning related to social due to its increased use in the collocation *global warming*. The distributions for other words are shown in the Appendix A.

6 Analysis of Amelioration and Pejoration

The process of mapping the BERT space to the Binder space not only improves the interpretability of LSC, as described in Section 5, but also facilitates the search for words corresponding to specific types of LSC. This section presents a case study to search for words that went through amelioration or pejoration, where amelioration means acquiring positive sentiment and pejoration means acquiring negative sentiment (Traugott, 2017). In addition, we evaluate the ability of our method to identify specific words that acquire a positive or negative meaning over time.

6.1 Known Words of Amelioration and Pejoration

Several pieces of literature have already reported examples of amelioration and pejoration. From these references, the sets of known words of amelioration and pejoration, $\mathcal{W}_{\text{ame}}$ and $\mathcal{W}_{\text{pej}}$ respectively, are extracted. Table 5 shows $\mathcal{W}_{\text{ame}}$ and $\mathcal{W}_{\text{pej}}$ with their references. Although both sets are small, they are used as ground truth to examine whether our method successfully identifies these words as amelioration or pejoration.

$\mathcal{W}_{\text{ame}}$	**hysteria** (Cook and Stevenson, 2010)
	brilliant, **fabulous**, **fantastic**, **spectacular** (Altakhaineh, 2018)
	terrific (de Wit, 2021)
$\mathcal{W}_{\text{pej}}$	**dynamic**, **synthesis** (Cook and Stevenson, 2010)
	abuse, **addiction**, **harassment**, **prejudice**, **trauma** (Haslam, 2016)
	terrible (Altakhaineh, 2018)
	awful (de Wit, 2021)

Table 5: Sets of known words of amelioration $\mathcal{W}_{\text{ame}}$ and pejoration $\mathcal{W}_{\text{pej}}$

6.2 Methods

First, we select the Binder features that are related to positive or negative meanings. Referring to Binder et al. (2016), the features related to positive meanings $\mathcal{I}_{\text{pos}}$ are defined as *Pleasant* and *Happy*, while the features related to negative meanings $\mathcal{I}_{\text{neg}}$ are defined as *Pain*, *Harm*, *Unpleasant*, *Sad*, *Angry*, *Disgusted*, and *Fearful*. Indeed, some of these features indicate that the LSC type of PC9 and PC3 in Table 4 are amelioration and pejoration, respectively. Furthermore, *Happy*, *Sad*, *Angry*, *Disgusted*, and *Fearful* are derived from the basic emotions proposed by Ekman (1992), which are closely related to emotion analysis (Plaza-del Arco et al., 2024) in the field of NLP.

Next, for each target word collected in Section 5.1, a score indicating the degree of positive or negative lexical semantic change (called LSC score in this paper) is calculated as follows:

$$\text{LSCS}(w, x) = \max_{i \in \mathcal{I}_x} \mathbf{v}_{\text{lsc}}(w)[i], \quad (6)$$

where $\mathcal{I}_x$ is either $\mathcal{I}_{\text{pos}}$ or $\mathcal{I}_{\text{neg}}$. That is, the maximum value of the positive (or negative) features in the LSC vector is employed as the LSC score. Our motivation behind this definition is that a word should be recognized as amelioration or pejoration if one of the features in $\mathcal{I}_{\text{pos}}$ or $\mathcal{I}_{\text{neg}}$ increases significantly.

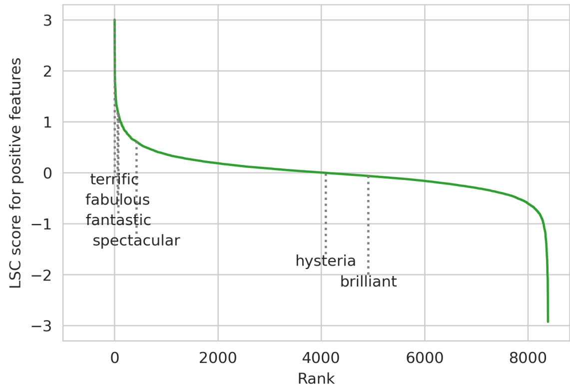


Figure 5: LSC scores for positive features

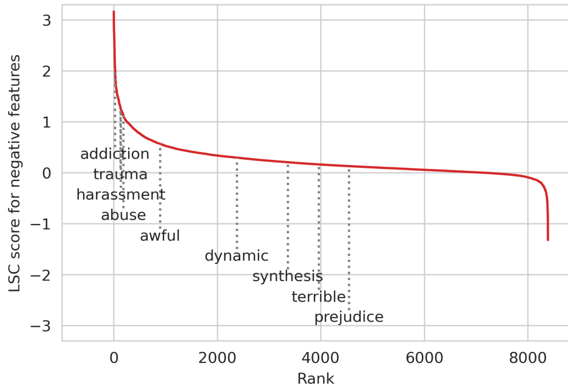


Figure 6: LSC scores for negative features

Finally, we sort all the target words in order of their LSC scores and verify whether the words in $\mathcal{W}_{\mathrm{ame}}$ or $\mathcal{W}_{\mathrm{pej}}$ are highly ranked.

While previous methods (Cook and Stevenson, 2010; Goworek and Dubossarsky, 2024) are specialized for detecting amelioration and pejoration, our approach can extend to identify words of other LSC types discovered in Section 5.

6.3 Results

Figures 5 and 6 show the LSC scores for positive features $\mathrm{LSCS}(w, \mathrm{pos})$ and negative features $\mathrm{LSCS}(w, \mathrm{neg})$, respectively. Words changing in a positive direction (i.e., the LSC score is greater than zero) account for about half of the total, while words changing in a negative direction account for about 75%. This indicates that words tend to change in a negative direction more than in a positive direction.

Figure 5 shows that the rank of most known words of amelioration in Table 5 are relatively high. In particular, *terrific* is ranked first. The OED and de Wit (2021) denote that *terrific* began to be used with a positive meaning in addition to a negative one in the late 19th century, and today it is mainly used with a positive meaning. On the other hand, the LSC scores for *hysteria* and *brilliant* are nearly zero. For *hysteria*, no positive meaning similar to those shown by Cook and Stevenson (2010) are found in the OED and examples in the CCOHA. For *brilliant*, according to the OED, this word originally meant *shining* and acquired the metaphorical meaning of *splendid* around 1739. This semantic shift was not captured because the LSC score is measured between periods of the 1910s and 2000s.

Figure 6 indicates that the meaning of all the known words of pejoration in Table 5 are shifted in a negative direction. The words *abuse*, *addiction*, *harassment*, and *trauma*, which are suggested by Haslam (2016), are ranked relatively high. According to Haslam (2016), as the meanings of these words expand, people become more sensitive to their negative connotations. On the other hand, the ranks of some words in $\mathcal{W}_{\mathrm{pej}}$ are low. For *prejudice*, the results are similar to those of Vylomova et al. (2019), and unlike other words in Haslam (2016), its meaning has not drastically shifted in a negative direction. For *dynamic* and *synthesis*, no negative meaning similar to those shown by Cook and Stevenson (2010) is found in the OED and examples in the CCOHA. For *terrible*, since this word has only negative meaning, it is unlikely that its meaning will change in a more negative direction.

To sum up, these results demonstrate the effectiveness of our method in the detection of amelioration and pejoration.

7 Conclusion

This study proposed a novel method to improve the interpretability of LSC by mapping the semantic space of the pre-trained language model to the neurobiological space. In the experiments designed to estimate the degree of LSC, our method demonstrated better performance than the baseline methods that did not map the semantic spaces. By leveraging the high interpretability of our method, we discovered interesting types of LSC that had not been identified previously. Additionally, in the detection of amelioration and pejoration, our method assigned appropriate LSC scores for words, which evaluated how their meanings changed positively or negatively. In the future, we plan to apply our method to detect words of other types of LSC.

Limitations

In this study, we analyzed several LSC types from the perspective of the Binder features. On the other hand, according to Traugott (2017), there are different types of LSC, such as metaphorization, metonymization, narrowing, and generalization. The method proposed in this paper might struggle to capture these LSC types because there is no clear correlation between the Binder features and them. Therefore, it is necessary to extend the current method or adopt new methods of representation (e.g., representing the meaning of a word in a sentence with box embeddings (Oda et al., 2024)).

In addition, it is necessary to increase the number of target words. In our method, words that are not included in the vocabulary of the tokenizer of pre-trained language models are outside the scope of the analysis, resulting in failure to capture the LSC of those words. Even when a word is split into multiple subwords, contextualized embeddings should be obtained, for example, by taking an average vector of the contextualized embeddings of these subwords (Montariol et al., 2021).

References

Taichi Aida and Danushka Bollegala. 2023. Swap and predict – predicting the semantic changes in words across corpora by context swapping. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7753–7772, Singapore. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2024. A semantic distance metric learning approach for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7570–7584, Bangkok, Thailand. Association for Computational Linguistics.

Taichi Aida and Danushka Bollegala. 2025. Investigating the contextualised word embedding dimensions specified for contextual and temporal semantic changes. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1413–1437, Abu Dhabi, UAE. Association for Computational Linguistics.

Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2020. CCOHA: Clean corpus of historical American English. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6958–6966, Marseille, France. European Language Resources Association.

Cindy Aloui, Carlos Ramisch, Alexis Nasr, and Lucie Barque. 2020. SLICE: Supersense-based lightweight interpretable contextual embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3357–3370, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abdel Rahman Mitib Altakhaineh. 2018. The semantic change of positive vs. negative adjectives in modern english. *Lingua Posnaniensis*, 60(2):25–38.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 380–389. JMLR.org.

Jeffrey R. Binder, Lisa L. Conant, Colin J. Humphries, Leonardo Fernandino, Stephen B. Simons, Mario Aguilar, and Rutvik H. Desai and. 2016. Toward a brain-based componential semantic representation. *Cognitive Neuropsychology*, 33(3-4):130–174. PMID: 27310469.

Lyle Campbell. 2020. *Historical Linguistics*. Edinburgh University Press, Edinburgh.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 590–602, Toronto, Canada. Association for Computational Linguistics.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC pretrained model for cross-lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585, Toronto, Canada. Association for Computational Linguistics.

Emmanuele Chersoni, Enrico Santus, Chu-Ren Huang, and Alessandro Lenci. 2021. Decoding word embeddings with brain-based semantic features. *Computational Linguistics*, 47(3):663–698.

Paul Cook and Suzanne Stevenson. 2010. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ilse de Wit. 2021. *A Terrific Paper: A Corpus Study of Amelioration and Pejoration in Adjectives Related to Fear*. Ph.D. thesis, Stockholm University, Faculty of Humanities, Department of English.

Marco Del Tredici, Raquel Fernández, and Gemma Boleda. 2019. Short-term meaning shift: A distributional exploration. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2069–2075, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Martin Emms and Arun Kumar Jayapal. 2016. Dynamic generative model for diachronic sense emergence detection. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1362–1373, Osaka, Japan. The COLING 2016 Organizing Committee.

Mariia Fedorova, Andrey Kutuzov, and Yves Scherrer. 2024. Definition generation for lexical semantic change detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5712–5724, Bangkok, Thailand. Association for Computational Linguistics.

Natalia Flechas Manrique, Wanqian Bao, Aurelie Herbelot, and Uri Hasson. 2023. Enhancing interpretability using human similarity judgements to prune word embeddings. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 169–179, Singapore. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Mario Giulianelli, Iris Luden, Raquel Fernandez, and Andrey Kutuzov. 2023. Interpretable word sense representations via definition generation: The case of semantic change analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3130–3148, Toronto, Canada. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.

Roksana Goworek and Haim Dubossarsky. 2024. Toward sentiment aware semantic change analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 350–357, St. Julian's, Malta. Association for Computational Linguistics.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

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

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908, Florence, Italy. Association for Computational Linguistics.

Seiichi Inoue, Mamoru Komachi, Toshinobu Ogiso, Hiroya Takamura, and Daichi Mochihashi. 2022. Infinite SCAN: An infinite model of diachronic semantic change. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1605–1616, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65, Baltimore, MD, USA. Association for Computational Linguistics.

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*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Severin Laicher, Sinan Kurtyigit, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. 2021. Explaining and improving BERT performance on lexical semantic change detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 192–202, Online. Association for Computational Linguistics.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021a. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yang Liu, Alan Medlar, and Dorota Glowacka. 2021b. Statistically significant detection of semantic shifts using contextual word embeddings. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 104–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020a. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020b. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 343–349, New York, NY, USA. Association for Computing Machinery.

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

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Kohei Oda, Kiyoaki Shirai, and Natthawut Kertkeidkachorn. 2024. Learning contextualized box embeddings with prototypical networks. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 1–12, Bangkok, Thailand. Association for Computational Linguistics.

Abhishek Panigrahi, Harsha Vardhan Simhadri, and Chiranjib Bhattacharyya. 2019. Word2Sense: Sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5692–5705, Florence, Italy. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

David Rother, Thomas Haider, and Steffen Eger. 2020. CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 187–193, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Anna Hätty, Marco Del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings* of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2019. Survey of computational approaches to lexical semantic change. *Preprint*, arXiv:1811.06278.

Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen, editors. 2021. *Computational approaches to semantic change*. Number 6 in Language Variation. Language Science Press, Berlin.

Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. 2017. Analyzing semantic change in Japanese loanwords. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1195–1204, Valencia, Spain. Association for Computational Linguistics.

Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. Can word sense distribution detect semantic changes of words? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3575–3590, Singapore. Association for Computational Linguistics.

Elizabeth Closs Traugott. 2017. Semantic change. *Oxford Research Encyclopedia of Linguistics*.

Jacob Turton, Robert Elliott Smith, and David Vinson. 2021. Deriving contextualised semantic features from BERT (and other transformer model) embeddings. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 248–262, Online. Association for Computational Linguistics.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 1–8, Marseille, France. European Language Resources Association.

Akira Utsumi. 2018. A neurobiologically motivated analysis of distributional semantic models. *Preprint*, arXiv:1802.01830.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

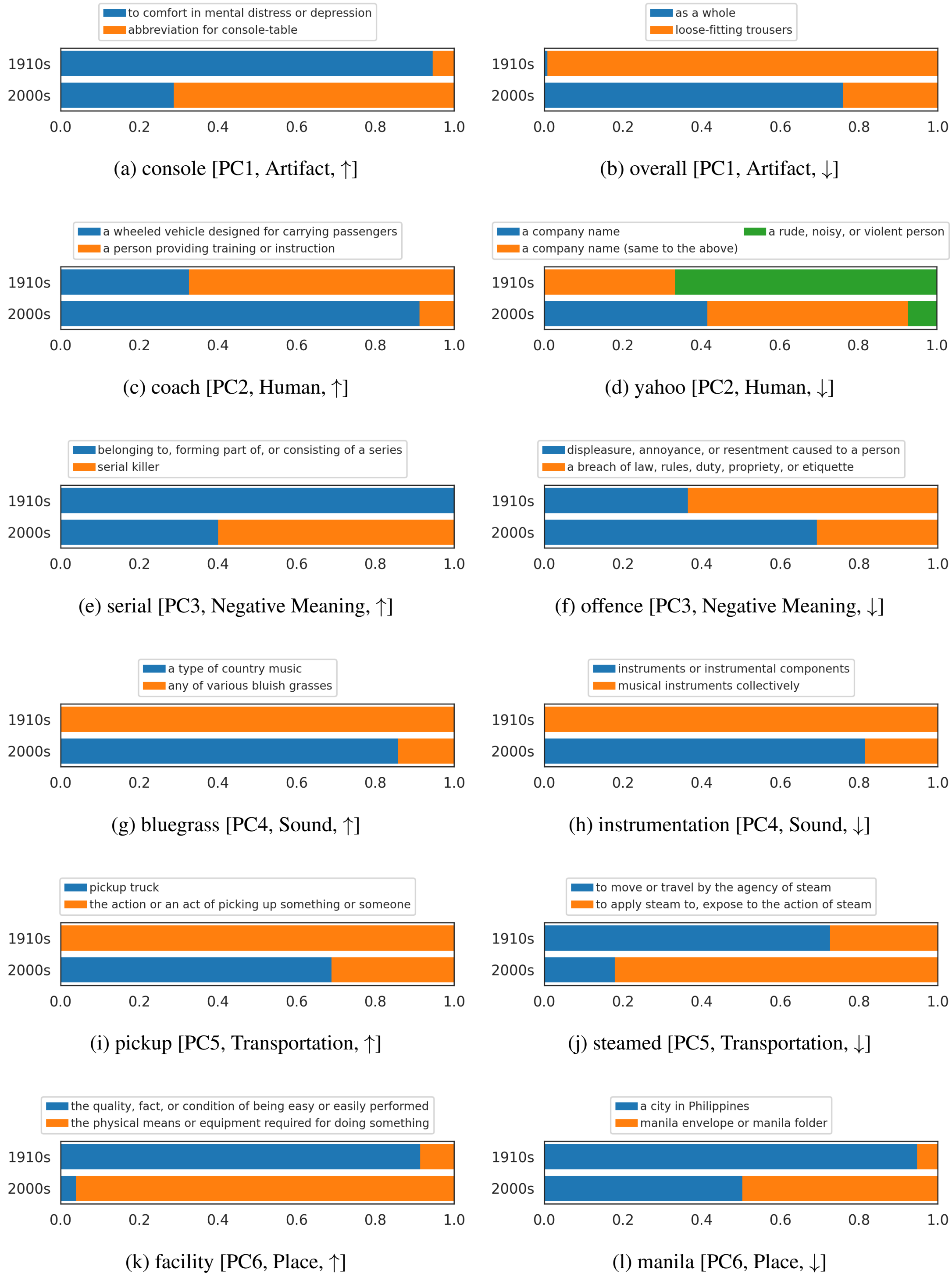
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ekaterina Vylomova, Sean Murphy, and Nicholas Haslam. 2019. Evaluation of semantic change of harm-related concepts in psychology. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Lütfi Kerem Şenel, İhsan Utlu, Furkan Şahinuç, Haldun M. Ozaktas, and Aykut Koç. 2020. Imparting interpretability to word embeddings while preserving semantic structure. *Natural Language Engineering*, 27(6):721–746.

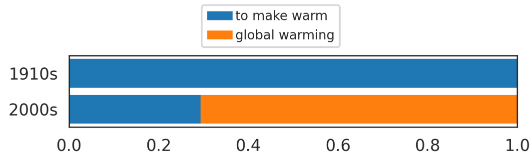
A Distributions of the usage types

The distributions of the usage types for several representative words in Table 4 are shown in Figures 7 and 8.

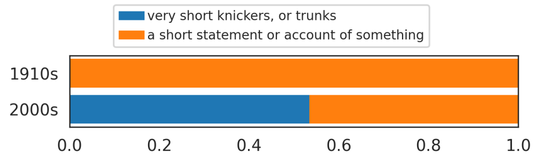


(a) console [PC1, Artifact, ↑]
(b) overall [PC1, Artifact, ↓]
(c) coach [PC2, Human, ↑]
(d) yahoo [PC2, Human, ↓]
(e) serial [PC3, Negative Meaning, ↑]
(f) offence [PC3, Negative Meaning, ↓]
(g) bluegrass [PC4, Sound, ↑]
(h) instrumentation [PC4, Sound, ↓]
(i) pickup [PC5, Transportation, ↑]
(j) steamed [PC5, Transportation, ↓]
(k) facility [PC6, Place, ↑]
(l) manila [PC6, Place, ↓]

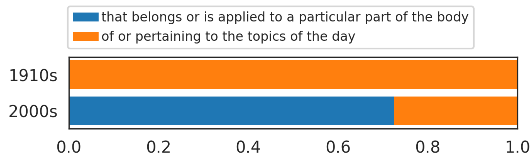
Figure 7: Distributions of usage types of representative words. These words are excerpted from Table 4. An ID of a principal component, an LSC type label, and an arrow indicating the direction of semantic change of each word are in parentheses.



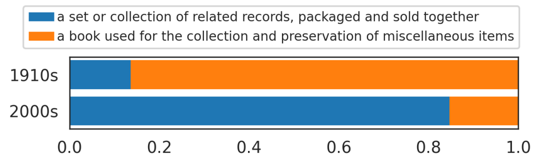
(m) warming [PC7, Social, ↑]



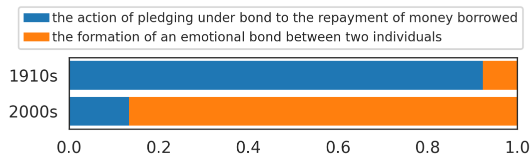
(n) briefs [PC7, Social, ↓]



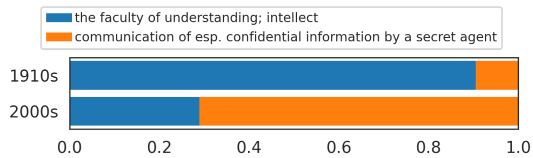
(o) topical [PC8, Familiar Thing, ↑]



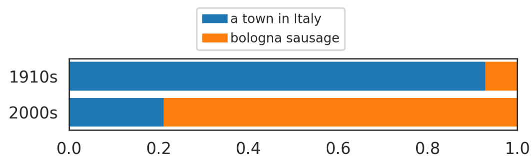
(p) album [PC8, Familiar Thing, ↓]



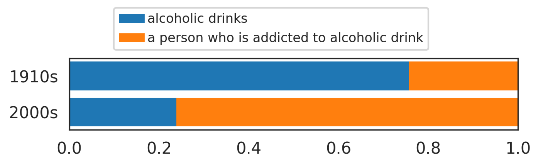
(q) bonding [PC9, Positive Meaning, ↑]



(r) intelligence [PC9, Positive Meaning, ↓]



(s) bologna [PC10, Food, ↑]



(t) alcoholic [PC10, Food, ↓]

Figure 8: Distributions of usage types of representative words (cont.)