Probing visualizations of neural word embeddings for lexicographic use

Ágoston Tóth[1], Esra Abdelzaher[2]

[1] University of Debrecen, Faculty of Humanities, Department of English Linguistics
[2] University of Debrecen, Institute of English and American Studies
University of Debrecen, Doctoral School of Linguistics
E-mail: toth.agoston@arts.unideb.hu, esra.abdelzaher@gmail.com

Abstract

Our study explores the possibility of using the distributional characteristics of headwords as exemplified in the online Oxford Learner's Dictionaries, captured by contextualized word embeddings and displayed in two dimensions to help lexicographers find sense categories, detect variations across senses and select potential example sentences. In addition to the dictionary examples, we added British National Corpus data that contained the headwords. BERT word embeddings were extracted for all occurrences of the headword, then two-dimensional representations of the resulting high-dimensional BERT embedding vectors were created using 4 algorithms: MDS, Isomap, Spectral and t-SNE. Clustering was assisted by k-means clustering and Silhouette scoring for different k values. Our investigation showed that Silhouette scores for k-means increased after dimension reduction; furthermore, spectral and t-SNE visualizations were associated with the most cohesive clusters. The highest Silhouette scores recommended a number of clusters different from the number of dictionary senses, but semantic and syntactic patterns were detectable across the recommended clusters.

Keywords: sense delineation; word embedding visualization; BERT

1. Introduction

Lexicography is open to incorporating advances in information technology, especially when a large amount of manual labour can be substituted. Consider how quickly concordancing became computerized, also the swift adaptation of database management systems to store lexicographic data, or the introduction of methods for quantitative corpus analysis, including those for detecting potential collocations via scoring first-order (syntagmatic) word co-occurrence patterns using t-score, MI-score, etc.

The idea that word distribution can be directly exploited for capturing meaning was pointed out by Firth (1957), who famously argued that the meaning of a word is distributed over the neighbouring words, or the company that words keep. Words may be distributionally similar (therefore, they appear in paradigmatic relations in their second-order co-occurrence patterns) for semantic and structural reasons; the presence of the semantic component is now being actively exploited in Natural Language Processing and Artificial Intelligence research. In what follows, we will refer to this area

of interest as Distributional Semantics (DS; cf. Lenci, 2008).

In the 2010s, the quick spread of connectionist language modelling and the eventual introduction of Large Language Models (LLMs) changed Distributional Semantics in its implementation, and expanded the range of applications in Natural Language Processing. Machine learning algorithms based on artificial neural networks get distributional data from large amounts of text while learning to solve distribution-related tasks (such as masked-word prediction, next-word prediction and context prediction). While doing so, they internally characterize the tokens of the text that they are processing; we call these internal characterizations *word embeddings*. The latest generation of LLMs, which includes the ELMo model (Peters et al., 2018), BERT (Devlin et al., 2019) and GPT (Radford et al., 2018), are designed to dynamically associate actual uses of tokens with their distributional features, giving us *contextualized* embeddings. It is reasonable to evaluate whether contextualized word embeddings can be used for identifying senses for lexicographic use, too.

Sense delineation presents a significant challenge to practicing lexicographers, given the complexity and fuzziness of meaning categories. Explaining the meaning of a simple word such as *dog* requires knowledge about multiple semantic fields including shape, movement and sound. Linguists have the means to discuss the complexity of the meaning of words and how they may overlap when sharing the same conceptual base or schematic structure (e.g. Langacker, 1999; Lakoff, 1987 and Fillmore & Atkins, 1992). Lexicographers, however, need to represent word meaning as a finite list of senses. In this regard, deducing word senses from corpus uses is very challenging. Using the target word as part of a name or sublanguage is likewise problematic for lexicographers. Lexicographers have to decide whether this is a different unpredictable sense that should be recorded in a dictionary or not. Moreover, non-standard word use always depends on deviation from the known use. However, the new use is not always salient for users, specifically if triggered by a combination of words rather than a single target word (Kilgarriff, 2007).

In this paper, we explore the possibility of employing BERT word embeddings as tools for identifying senses of words as they appear in dictionary examples and also in additional corpus sentences. Section 2 of this paper discusses related work in the literature. Section 3 presents the methodology of the current research from data collection, through producing 2-dimensional visualizations that may assist lexicographic work, to the examination of the clusters. Section 4 has the qualitative analysis of the visualizations for the four words that we have selected for this analysis. Our concluding remarks are presented in Section 5, where we also discuss the limitations of our research.

2. Related work

Rychlý & Kilgarriff (2007) offered a DS method for building distributional thesauri. They used a corpus of lemmatized and parsed language to gather information about

how words are used in context, including the grammatical relations between a target word and other (context) words in sentences. The method then identifies other words that share similar contexts. This function is also available in the Sketch Engine, where "Sketch differences" rely on lexical collocates and grammatical relations in the contexts to show how (dis)similar two words are (Kilgarriff et al., 2014). This type of information has been useful in unveiling word senses that are not present in dictionaries (see, for instance, Abdelzaher & Tóth, 2020). The "Sketch differences" tool does not use contextualized word embeddings.

Jatowta, Tahmasebi & Borin (2021) give a review of the literature that tracks meaning change in a diachronic setting using distributional data of words, and tackle the question of visualization, too. The paper illustrates that even static embeddings can help us compare different states of the language if we generate snapshots for the states under scrutiny, generate static embeddings for them and compare these embeddings. Unfortunately, static embeddings contain a mix of all senses, all usages of the given word, so they cannot directly help the sense delineation process. The possibility of using contextualized word embeddings is pointed out by the authors as a possible future direction.

Montes & Heylen (2022) visualize distributional semantic data for testing different word embedding parameter sets (which is common practice with static "count-type" embeddings) and also for checking the distributional properties of the word under scrutiny – the Dutch word *heffen* with 2 senses. Their study is presented in the context of cognitive linguistics. In our present paper, we utilize a single, pre-trained distributional model that implements a modern contextualized word embedding type designed to collect token-level distributional information in a context-sensitive way; the parameters that we test are related to the visualization step rather than distribution modelling, and our focus is on sense delineation within the context of lexicography.

In our work, we use BERT word embeddings (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019), which is a well-established contextualized embedding type in Natural Language Processing. BERT is based on the Transformer architecture (Vaswani et al., 2017). The model learns to predict a masked word in a sentence and to decide if two sentences appeared sequentially in the training corpus. As a contextualized model, BERT captures the distributional properties of actual uses of words (more precisely, those of tokens in its vocabulary) in given contexts. Outside of the field of lexicography, contextualized word embeddings have been proven to form distinct clusters corresponding to different word senses in Wiedemann et al. (2019).

3. Methods

3.1 Data collection

In our analysis, we created two-dimensional (2D) visualizations of BERT embeddings for instances of four headwords: *full*, *mouth*, *risk* and *sound*, as exemplified in dictionary example sentences and found in the British National Corpus.

The professionally selected and edited *dictionary* examples were taken from the online *Oxford Learner's Dictionaries* at http://www.oxfordlearnersdictionaries.com (OD). We took all examples (including the "Extra Examples") of the selected headwords in all senses, but we had to discard those examples that contained an inflected form of the headword, as inflected forms are treated as different BERT tokens (which may get related in their representations, but the analysis of the relation between the embeddings of headwords and inflected forms is beyond the scope of this paper) or, in some cases, sequences of tokens. Hornby's *Idiomatic and Syntactic English Dictionary* (Hornby, 1948), which is known for its inclusion of syntactic information and its focus on word complementation, is part of OD's heritage, which may be reflected in the example sentences OD provides for each word sense. For this reason, different syntactic patterns corresponding to different senses are expected to stand out in the visualized representations.

The additional *corpus* sentences (1000 for each headword) were taken from the British National Corpus (BNC) available via http://www.sketchengine.eu. We used the sentence concordancer option, looked up the word, shuffled the output and exported the data. We did not filter for part of speech. While BNC may not be the most extensive or most up-to-date corpus of English, it is a balanced representation of British English (Leech, 1992). We collected examples that contained the exact headword.

3.2 Creating BERT embeddings

We produced contextualized word embeddings for the headwords in the dictionary example sentences and corpus examples. The embeddings were created using the Huggingface BERT libraries (https://huggingface.co). We relied on a pre-trained BERT model (*bert-large-uncased,* https://huggingface.co/bert-large-uncased) and the corresponding *bert-large-uncased* tokenizer from Huggingface. The BERT-large model contains 336 million trained parameters with 24 layers and 16 attention heads. We did not fine-tune the network, as we wanted to visualize pure distributional data acquired for the standard BERT learning goals. The resulting word embeddings were vectors that contained 1024 floating point numbers for each use of the given headword in the dictionary examples and corpus sentences; we used the embedding developed in the last layer of BERT in the position of the target word. According to the distributional

hypothesis, more similar uses of the target words are in closer proximity to one another when we visualize distributional feature vectors in the resulting 1024-dimensional space.

3.3 Dimension reduction

We used manifold learning algorithms for dimension reduction from 1024 to 2 dimensions as they are capable of preserving the underlying structure of the data.

We employed four algorithms: Multidimensional Scaling (MDS), Isomap, Spectral and t-SNE. MDS is a linear method, which is computationally efficient, while the three non-linear methods should be able to learn more complex relationships between the data dimensions.

MDS creates a low-dimensional representation by minimizing the difference between distances of data point pairs in the high-dimensional space and pairwise distances in the low-dimensional space. The main contributions to the field of MDS are reviewed in Groenen & Borg (2014).

Isomap (Tenenbaum, de Silva & Langford, 2000) is based on graph theory. It uses geodesic distance, which is a path between two points on a surface – rather than along a straight line. The Isomap graph is created by connecting neighbouring points and computing the geodesic distance between each pair of points. The algorithm uses MDS to embed the data into a low-dimensional space preserving the pairwise geodesic distances.

Spectral clustering employs the graph Laplacian to encode the similarity between data points. The top eigenvectors of the Laplacian matrix are considered to capture the global structure of the data. Spectral embedding is known to be able to capture non-linear structures and different types of relationships. For details, see Ng, Jordan & Weiss (2002).

Finally, t-SNE (van der Maaten & Hinton, 2008) is a non-linear method that constructs a probability distribution over pairs of high-dimensional data points and a similar distribution over pairs of low-dimensional points, and it minimizes the difference between these two distributions using gradient descent in an iterative fashion. t-SNE is considered very effective at preserving the local structure of data at the expense of non-local structure.

t-SNE is often used in current Natural Language Processing research for dimension reduction. It is the infrequent use of the remaining three methods that led us to test the possibility of utilizing them for the task at hand. We suppose that lexicographers carrying out the manual evaluation of corpus data, and looking for – otherwise hidden – second-order co-occurrence patterns, would benefit from getting access to multiple methods to work with. Compare it to the range of tools we can use for detecting

potential collocates (and, in general, first-order co-occurrence patterns): t-core, MI-score, etc.

We used a free tool, the Orange Data Mining toolkit (Demsar et al., 2013; https://orangedatamining.com) for converting the 1024D token embeddings to 2D using the above manifold learning algorithms, and also for visualization of the 2D outputs as scatterplots. Figures 2, 3, 5, 6, 7, 8 and 9 of this paper were prepared using this program. The interactive scatterplots that you have access to while using the toolkit also offer zoom functionality and can show or hide sentences as data labels. These interactive services, which are not shown in this study, made an important contribution to our work. Note, however, that the Orange toolkit is not designed to be a "lexicographer's workbench".

3.4 *k*-means analysis of the clusters using Silhouette scores

In addition to visual observation of the low-dimensional representations, we also studied the original high-dimensional feature space and its 2D representations using *k*-means clustering with additional Silhouette scoring for selecting *k*.

K-means clustering is commonly used for grouping data points into clusters automatically, based on their similarity to each other. In our case, k centroids are initially selected using the *k*-means++ algorithm (Ostrovsky et al. 2006). Then data points are assigned to the closest centroids based on squared Euclidean distances. After this assignment step, an update step is carried out, which recalculates the centroids to optimize the overall result of the clustering. In our experiment, we allowed for a maximum of 5000 iterations over the assignment and update steps. The algorithm is sensitive to the initial selection of the centroids (even with the *k*-means++ initial centroids); therefore, 20 reruns were performed, and the run with the lowest within-cluster error (lowest sum of squares) was kept.

The selection of the number of the clusters is of special importance in our case. It runs parallel to the lexicographic task of sense delineation, which involves drawing borderlines between senses, polysemous and homonomous, where polysemous senses are related in their meaning by definition. The lexicographical task of splitting and lumping senses is known to be challenging, and it is not automatized. In our exploratory research, we took OD's senses as reference points, but we also wanted to know the number of clusters that BERT data (raw and 2D-converted) naturally exhibited. Therefore, we used Silhouette scoring (Rousseeuw, 1987) of different k values in *k*-means analysis. Silhouette scoring is a measure of how well data points fit into their clusters, and it "shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters" (ibid.). A higher score indicates better clustering.

We carried out *k*-means clustering and calculated the Silhouette scores using the Orange Data Mining toolkit. We did not perform added quantitative evaluation of the clusters

(using Rand index or V-measure, for instance) in addition to what we have access to in the toolkit. Quantitative and qualitative analyses of the resulting plots are provided in the next section.

4. Results

4.1 Silhouette scores and *k*-means clusters before and after dimension reduction

Silhouette scores increased for all words after dimension reduction. In most cases, the number of clusters (C) was similar before and after dimension reduction and for the different visualization methods. However, for *risk*, the number of the suggested best clusters based on the 1024D distributional representations differed considerably from that recommended after t-SNE visualization. Figure 1 shows the Silhouette scores for different *k*-means clusters before and after the dimension reduction of the distributional representations of *risk*.

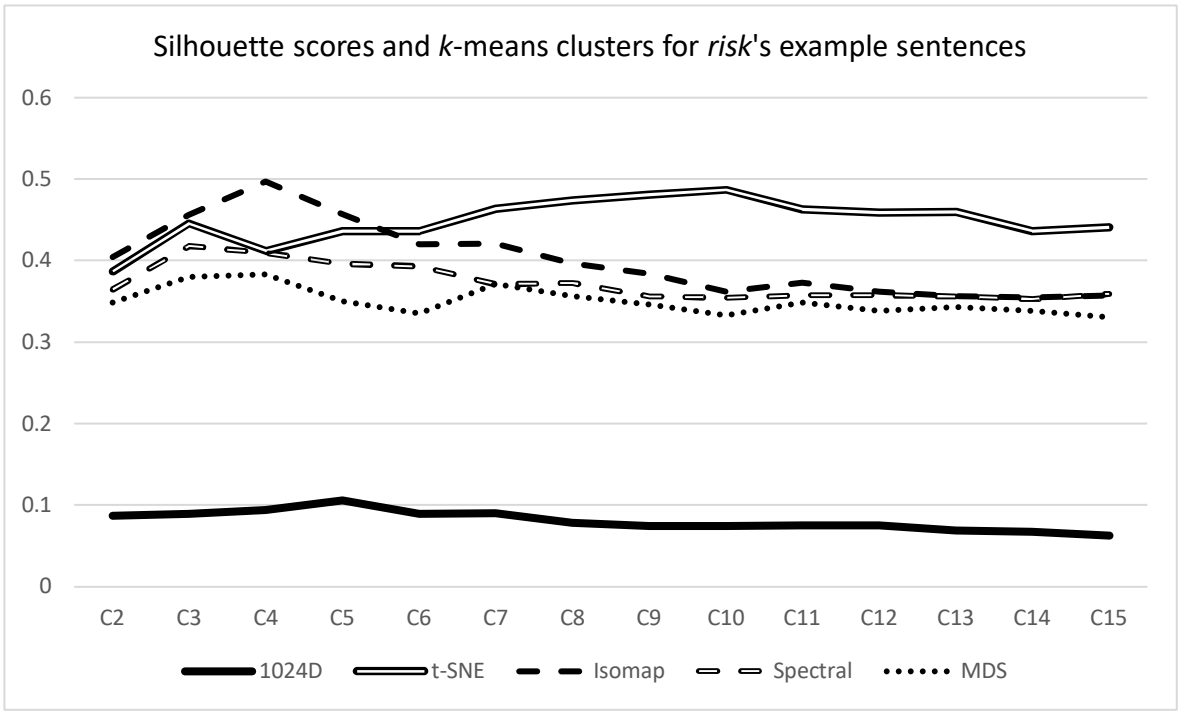

Figure 1: Silhouette scores for 2-15 clusters of *risk* before and after dimension reduction

The Silhouette scores for 2-15 clusters based on the 1024 dimensions represent an almost linear line on the chart without any significant peaks, at a consistently low value. On the contrary, for the t-SNE visualization, there is an increase in the Silhouette score for cluster three (0.456) and cluster ten (0.487). The best Silhouette score is associated with four clusters based on the Isomap visualization (0.497).

Before dimension reduction, the suggested five clusters hardly reflected any patterns. Figure 2 visualizes the box plot of the *k*-means clusters and a sample of the sentences in each cluster based on the 1024D representation of *risk*. Whereas the BNC sentences were distributed across the five clusters, the verbal senses of *risk* clustered together in C5. However, the same cluster usually contained heterogeneous sets of the uses of *risk*. C5 included the verbal senses of *risk* as recorded in the OD sentences and also had some of the nominal senses. C1 included only the nominal uses, but several contexts were present in the cluster. Medical risk was dominant in C1, but instances of *risk* in statistical and economic contexts appeared towards the end of the cluster. C2 was mostly associated with financial risks but also included several health-related risks towards the end of the cluster. Sentences in C3 referred to social, environmental, economic and medical risks. Sentences in C4 generally referred to risky situations without specification (at the top of the cluster) and associated *risk* with business loss and body injuries, among others.

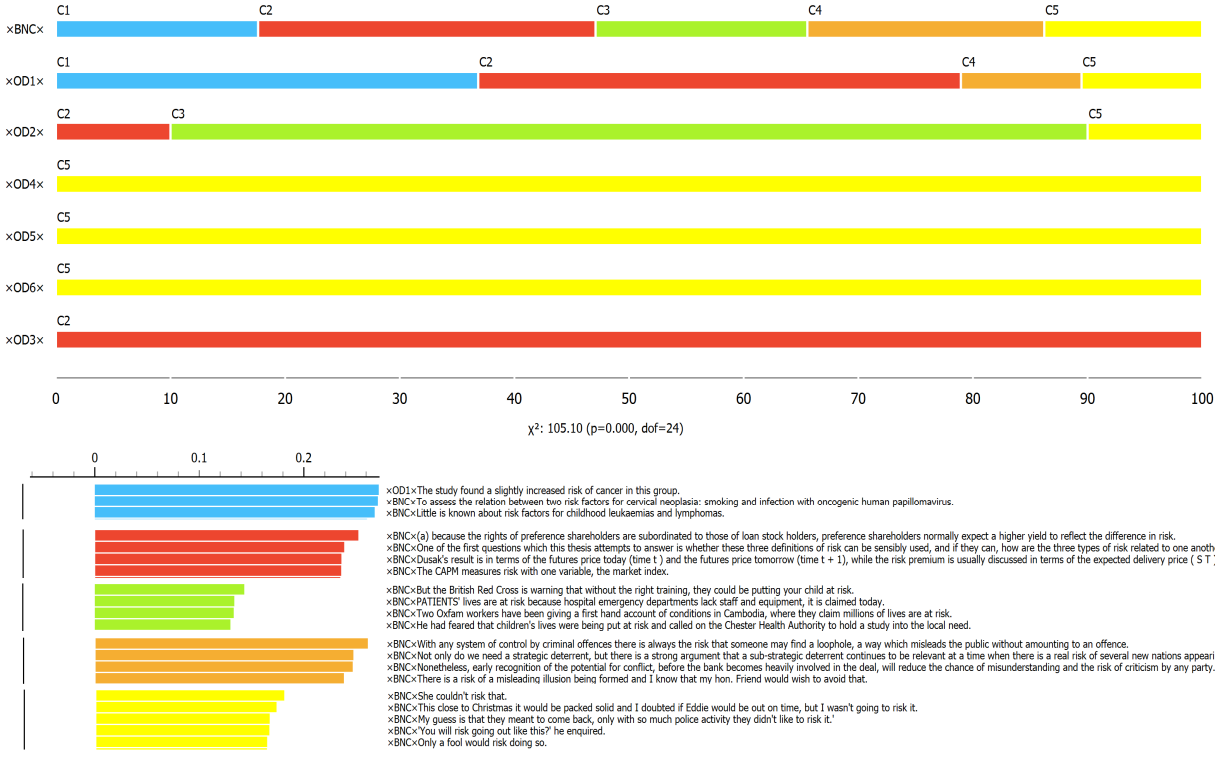

Figure 2: *k*-means clusters and sample sentences for *risk* in 1024D space

The increased Silhouette scores after the dimension reduction were reflected in the sentences grouped in each cluster. The suggested ten clusters based on the t-SNE visualization showed semantic and syntactic patterns shared among most of the sentences in a cluster. First, the verbal senses of *risk* clustered in C3 with verbal uses from the BNC, without nominal senses from OD in the cluster. Second, patterns, such as V_{be} *risk to* NP in C1, *increase(d)/reduce(d)/ high/ low risk of* NP in C2, *risk (of)+ing* and *risk+that*+clause in C4, started to appear in the clusters frequently. Third, compounds such as adj+*risk*+N were most frequent in C6, whereas collocates such as

at risk distinguished the sentences in C7. Fourth, sentences referring to health-related risks were primarily placed in C2, whereas business and financial risks dominated C5. Figure 3 displays the box plot of the *k*-means clusters and a sample of the sentences with *risk* after t-SNE visualization.

Figure 3: t-SNE-based *k*-means clusters and sample sentences for *risk*

Unlike the case of *risk*, the differences in the *k*-means clusters were minor for *mouth*. Figure 4 shows the Silhouette scores for *mouth* before and after dimension reduction. The Silhouette scores for different k values for the MDS visualization are almost similar, and they are considerably low. The best Silhouette score was 0.112 for two clusters before dimension reduction. After dimension reduction, the four visualization methods suggested three clusters as the best categorization of the five OD senses of *mouth* (i.e., part of the face, a person needing food, of a river, opening or entrance and way of speaking). The Silhouette score was best for the Spectral-based clusters (0.577), followed by Isomap (0.559), t-SNE (0.481) and MDS (0.375).

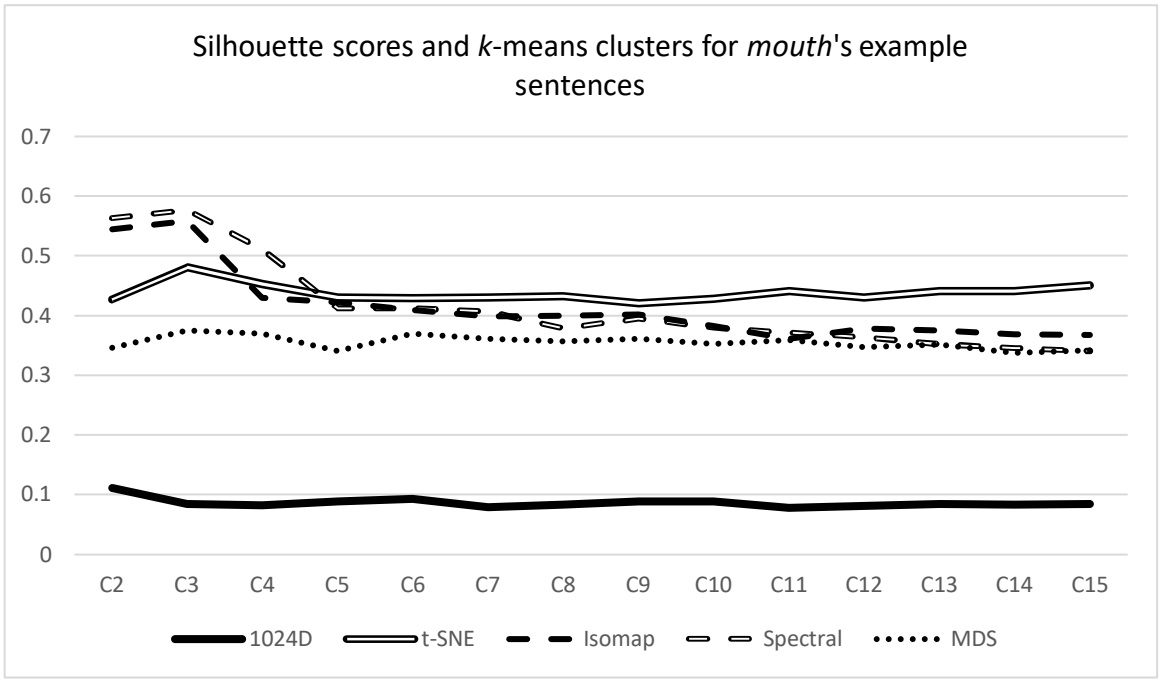

Figure 4: Silhouette scores for 2-15 k-means clusters for *mouth*'s example sentences

The remaining part of this section explores the sentences in the suggested two clusters based on the 1024D distributional representations and in the three clusters suggested based on the Spectral representation. Figure 5 shows the box plot of the k-means clusters for *mouth* in 1024D and the Silhouette plot of a sample of the sentences in the two clusters. As visualized, all OD senses are clustered in a single category, whereas a group of BNC sentences form a distinctive cluster. The first cluster contained a diaspora of heterogeneous sentences, and the second cluster mostly had sentences in which *mouth* was used in a romantic fiction genre. The literal sense of *mouth* (part of face), the metaphoric sense (opening) and the metonymic sense (way of speaking) appear in the same cluster.

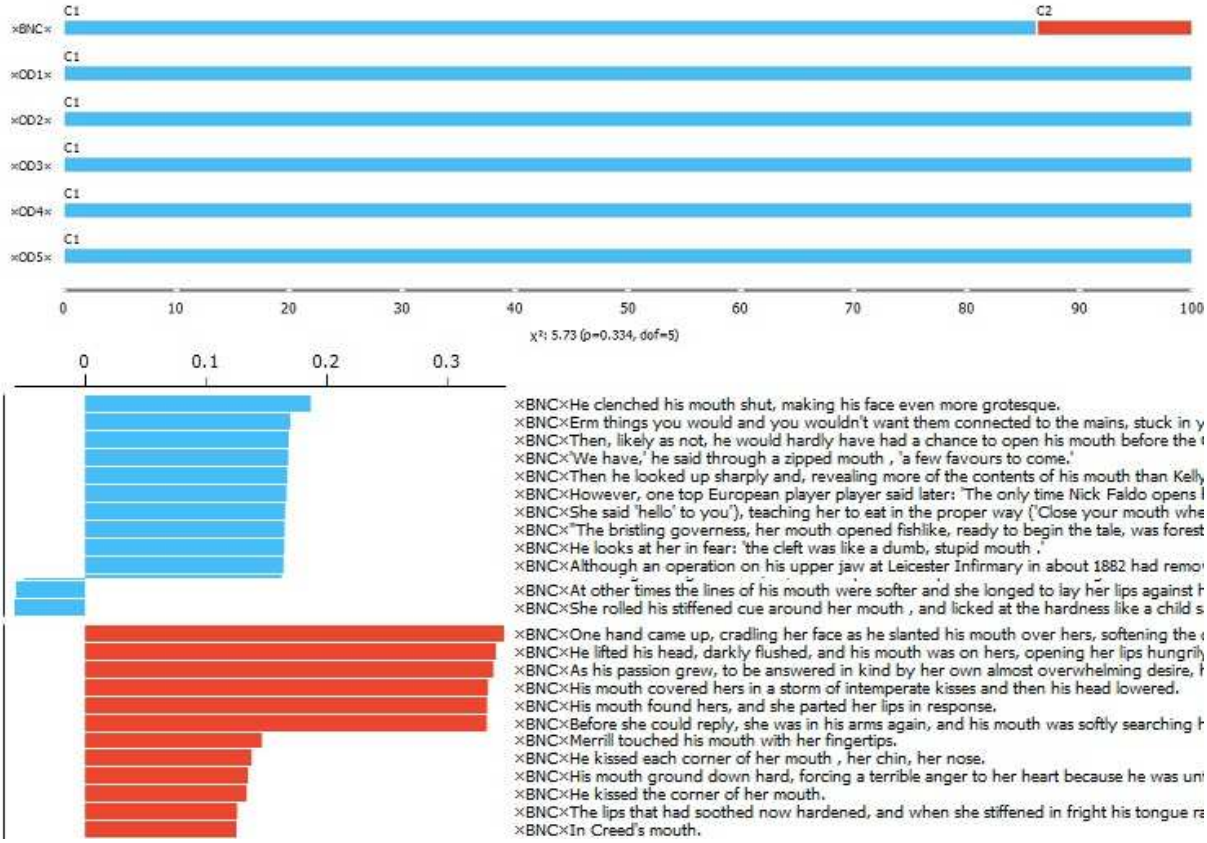

Figure 5: *k*-means clusters and sample sentences of *mouth* before dimension reductions

After dimension reduction, Spectral visualization showed the best Silhouette scores (0.577) for three clusters. The first cluster contained most of the senses of *mouth* (senses 2, 3, 4 and 5 in OD and some sentences from sense 1) and most BNC examples. Cluster two included the same romance-related uses of *mouth*, which clustered likewise before dimension reduction. However, a new category appeared and separated the uses of *mouth* to make facial expressions from other senses. The newly introduced cluster grouped sentences from OD's sense 1 and BNC examples.

4.2 Silhouette scores and *k*-means clusters: two perspectives

This section compares the best *k*-means clusters recommended by the Silhouette scoring to *k*-means clusters with *k* set to the number of dictionary senses. For *mouth*, the recommended clusters after using the four visualization methods were three as mentioned in the previous section (C3: making facial expressions, C2: romance-related sense, and C1: all other senses). We had five OD dictionary senses for *mouth*. Preselecting the number of clusters to five slightly improved the sub-clusters of the sentences, but it did not correspond to the dictionary senses. The three categories of *mouth* in romantic contexts, speaking and making facial expressions stood out again, although the literal use of the mouth to speak and the metonymic use as a way of speaking overlapped in clusters 1 and 5. The two added clusters contained a diaspora

of uses. For instance, cluster 1 included sentences referring to *mouth* in a medical context, as a way of speaking and with reference to eating and drinking. Cluster 5 grouped the metaphoric uses of *mouth* as 'mouth of a river' or 'entrance of a cave' with the literal uses of *mouth* in speaking. Figure 6 shows some of the similarity patterns in the sentences based on Spectral visualization of 5 *k*-means clusters.

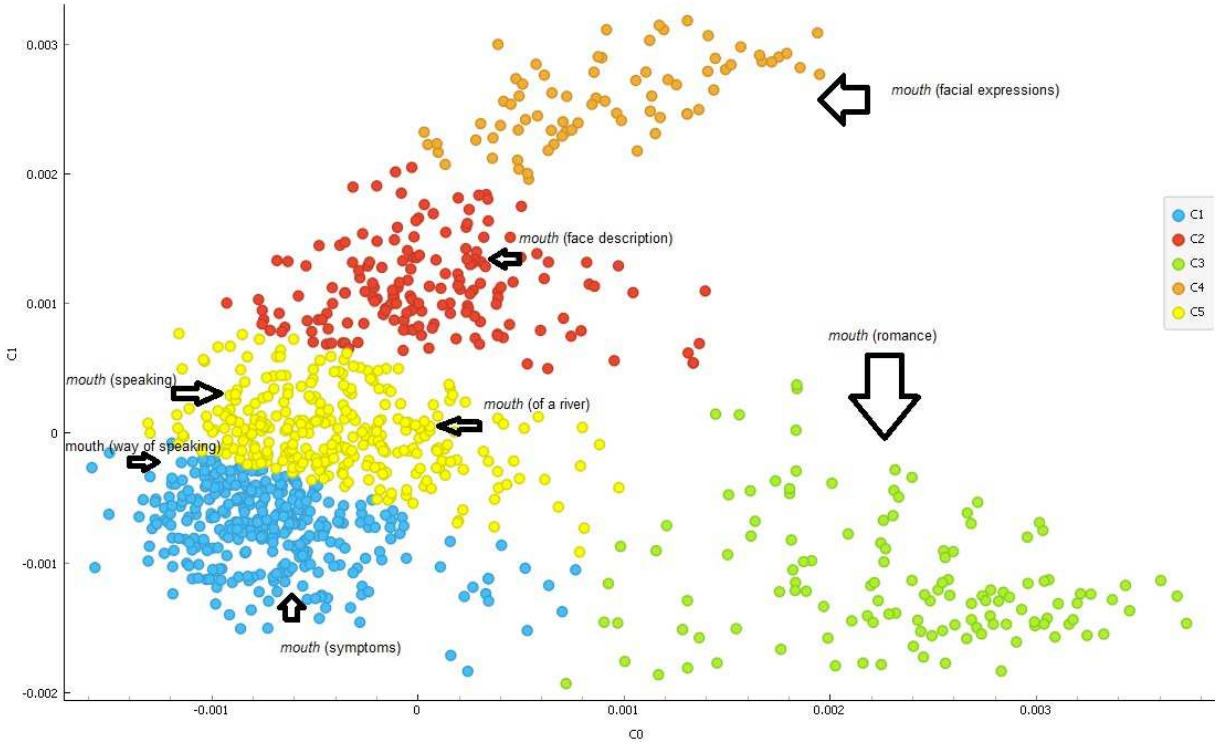

Figure 6: Scatter plot of 5 *k*-means cluster based on Spectral visualization of *mouth*'s sentences (colours indicate different *k*-means clusters as shown in the chart legend)

The same applies to *full*, which has 11 dictionary senses in the current study. However, before and after dimension reduction, the best Silhouette scores recommend two or three clusters for all the sentences of *full*. After manually setting the *k*-means clusters to 11, sentences in the clusters did not reflect the dictionary sense delineation. On the contrary, the same cluster contained semantically and syntactically dissimilar sentences whereas similar sentences overlapped in different clusters. As illustrated in figure 7, sentences expressing the literal and metaphoric senses of *full* as 'having a lot' appeared in four neighbouring clusters with no explicit patterns separating or joining them. In addition, the pattern *full* + noun which denotes 'complete' was frequent in two different categories.

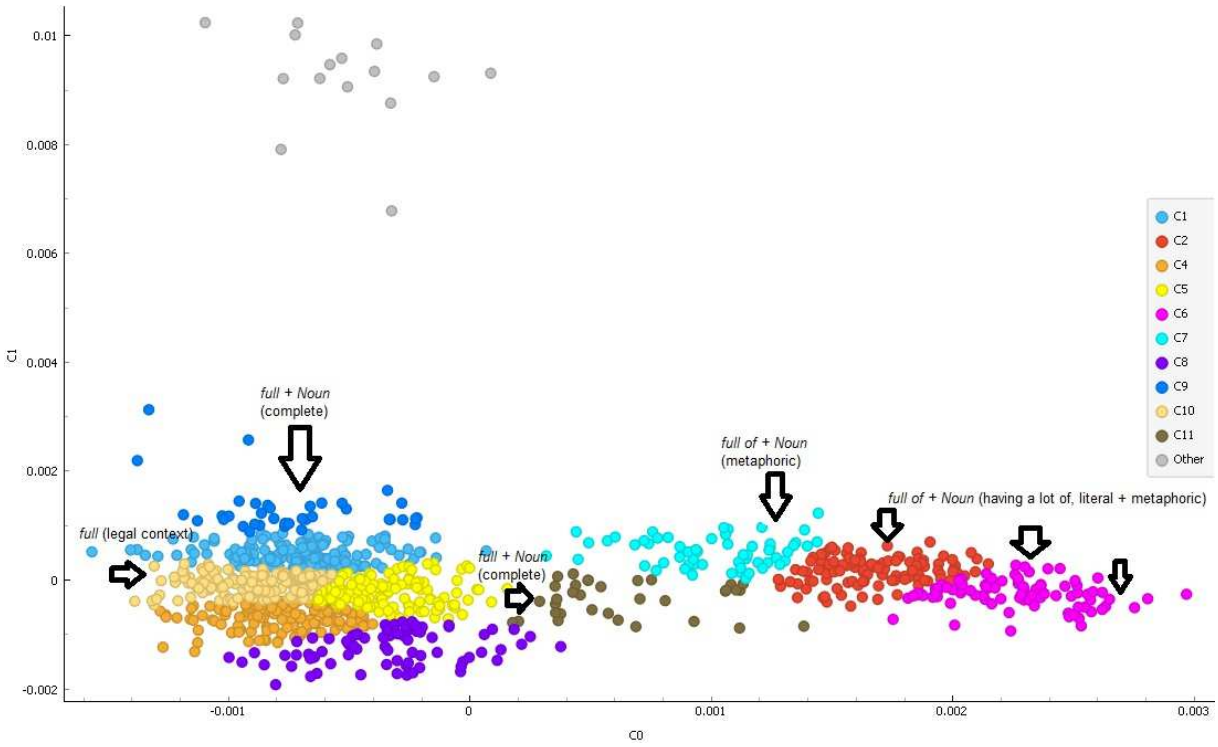

Figure 7: Spectral-based scatter plot of sentences with *full* in 11 pre-set *k*-means clusters

It is evident from the four case studies, investigated in this research, that pre-setting the number of clusters to match dictionary senses will not be helpful. However, depending on the automatically calculated highest Silhouette scores may be a better reflection of the patterns of use and, accordingly, of word senses, too, in or outside lexicographical contexts.

4.3 Comparing different visualization methods

Spectral, t-SNE and Isomap showed the best Silhouette scores for all words, unlike MDS. Figure 8 shows the four visualizations of the sentences of *sound* in a 2D space. Sentences are sporadically distributed all over the space with MDS, even if they instantiate the same sense. On the contrary, the visualized spaces created by Spectral, t-SNE and Isomap cluster the sentences closer to each other in major classes based on the part of speech. Sense categories are more salient in the t-SNE visualization of the examples of *sound*. First, the different parts of speech formed distinctive clusters all over the 2D space. Second, dissimilar senses belonging to the same POS appeared in different clusters. For instance, the nominal sense of *sound* as a passage of water appeared in a distinctive cluster other than the phonetics-, music- and television-related senses. Also, the verbal senses of *sound* as 'give impression' versus 'make a sound' appeared in two clusters with considerable distance between them. The similar nominal and verbal senses of *sound* as 'an impression' and 'give impression' formed close, but separate clusters.

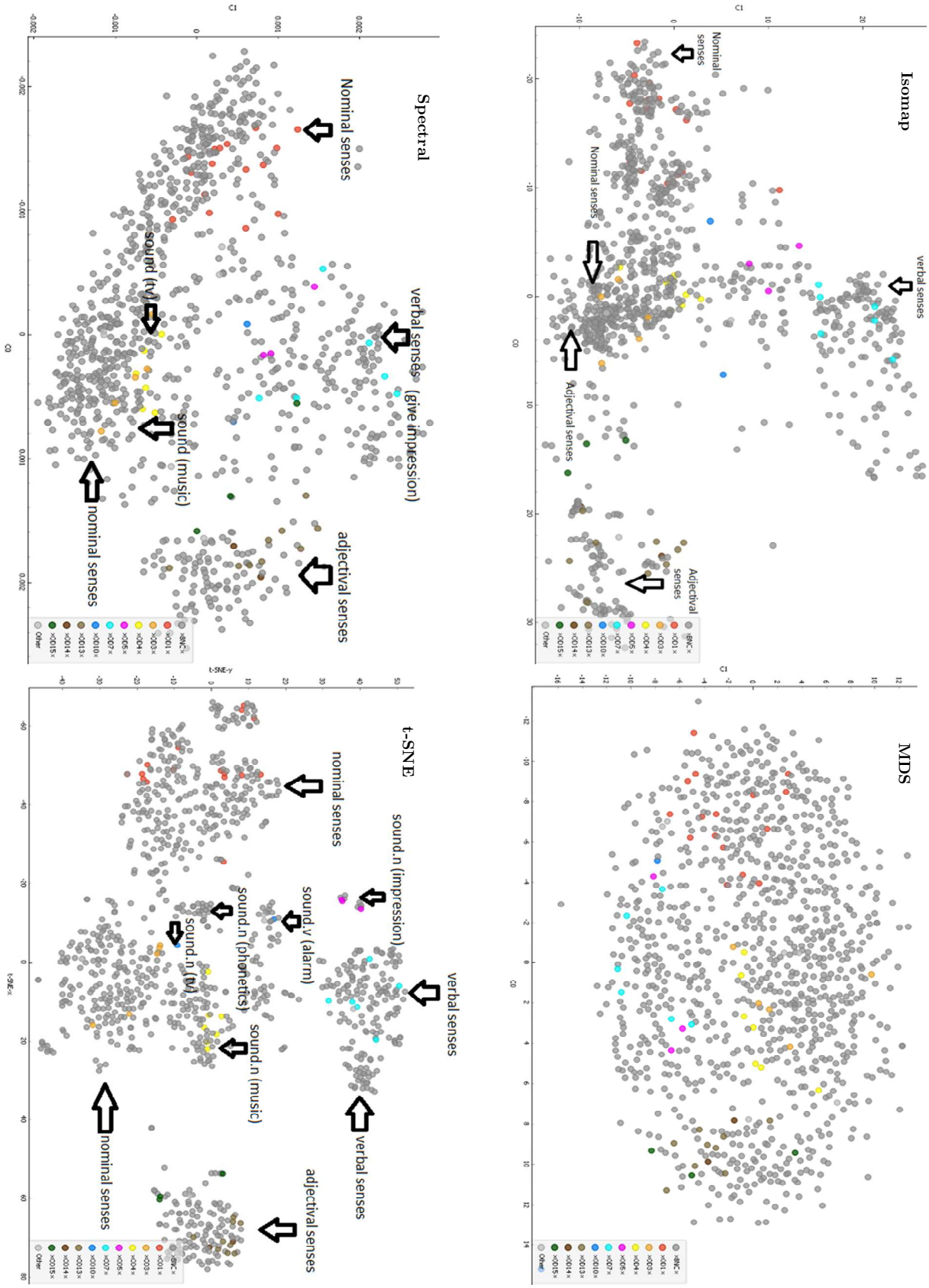

Figure 8: Four 2D visualizations of *sound*'s example sentences

In the initial phase of our experiments, manual parameter tuning was carried out based on Silhouette scores and also on the qualitative features of the resulting clusters, typically with one or two words. The parameter sets that we settled with for the different dimension reduction methods are shown in Table 1.

Dimension reduction	**Settings**
t-SNE	perplexity = 20
	distance = Eucledian
	initialization = PCA
	max. iterations = 3000
	learning rate = 200
MDS	initialization = PCA
	max. iterations = 5000
Isomap	neighbours = 20
Spectral	affinity = RBF kernel

Table 1: Parameter choices for the dimension reduction methods

We do not argue, however, that a single parameter set will cover all usage scenarios, all words of interest, all corpus sizes, etc. Instead, we recommend that the user should be given choices and the opportunity to find the most useful methods and settings. The t-SNE algorithm, for instance, is notoriously sensitive to the perplexity parameter, which balances the effect of local vs. distant neighbours on the resulting low-dimensional representation. We tried different values, and, in addition, we also explored different distance metrics, including Euclidean, Manhattan and Chebychev. Whereas the number of recommended clusters remained almost the same for all words, the Silhouette scores changed slightly. The best scores were mainly associated with the Euclidean metric and perplexity set as 20. Table 2 shows the suggested cluster numbers for *sound* corresponding to several t-SNE settings.

Distance metric	**Perplexity**	**Clusters**	**Silhouette Scores**
Euclidean	10	4	0.574
Euclidean	20	4	0.591
Euclidean	30	4	0.589
Manhattan	10	4	0.572
Manhattan	20	4	0.591
Manhattan	30	4	0.582
Chebychev	10	4	0.552
Chebychev	20	4	0.557
Chebychev	30	4	0.546

Table 2: The suggested clusters and Silhouette scores for sound in different t-SNE settings

Importantly, changing the parameters did not influence the inclusion of the OD sentences in the clusters or their overall position in the charts. The adjectival senses remained in the same cluster (C1) and appeared together on the t-SNE charts. Also,

the verbal and nominal senses of *sound* as 'to give an impression' and 'the idea or impression' were close to each other on the charts and formed a single cluster (C3). The nominal senses of *sound* with reference to phonetics, as a 'passage of water' and as 'audible signals' formed sub-clusters in cluster two (C2). The fourth cluster contained the verbal and nominal senses of *sound* as 'something you hear' and 'produce a sound'.

Changing the affinity measures for the Spectral algorithm had a considerable influence on the results. For *mouth, risk* and *sound*, the nearest neighbour affinity retrieved better results than RBF kernel. It was the opposite for the word *full,* however. Table 3 depicts the suggested clusters for all words using RBF kernel and nearest neighbour in the Spectral algorithm.

Word	**Affinity**	**Clusters**	**Silhouette score**
Full	RBF kernel	2	0.838
Full	Nearest neighbour	3	0.601
Mouth	RBF kernel	3	0.577
Mouth	Nearest neighbour	3	0.775
Risk	RBF kernel	3	0.418
Risk	Nearest neighbour	4	0.517
Sound	RBF kernel	4	0.529
Sound	Nearest neighbour	3	0.730

Table 3: The suggested clusters and Silhouette scores based on Spectral's affinity measures

Let us point out, however, that while the Silhouette scores increased with the nearest neighbour affinity, the homogeneity of the classes decreased in most cases. Figure 9 shows the distribution of the sentences with *mouth* over the Spectral space using the nearest neighbour measure. The cohesion of the clusters is evident, and the distance between some uses (e.g. 'using the mouth to make facial expressions' and 'reference to the mouth in face description') is noticeable. However, the overlap between the example sentences shows the heterogeneity of the sentences that form cohesive clusters. The figurative use of *mouth* as 'an opening of a hole or cave', the collocation *mouth open* with reference to surprise and *mouth* in relation to the medical field overlapped in the same cluster. Also, a mixture of literal and metaphoric uses of *mouth* and *open* were merged in the same cluster.

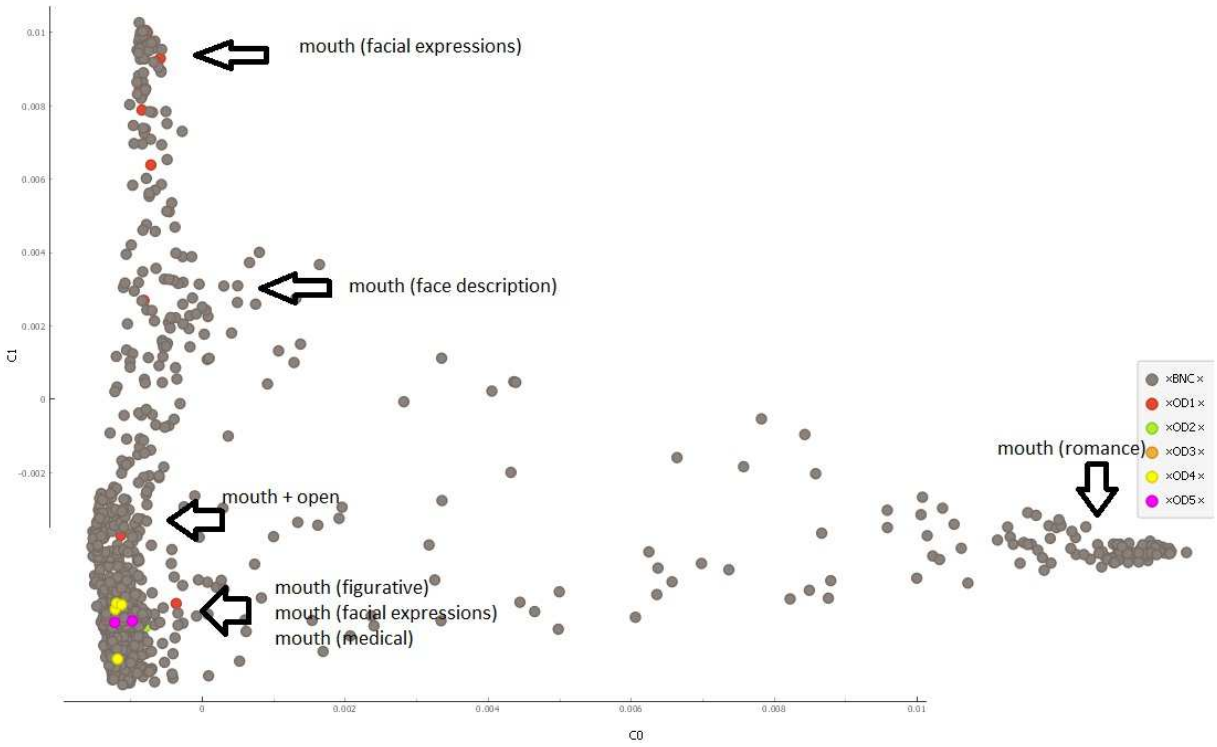

Figure 9: Spectral visualization using nearest neighbour measure for *mouth* sentences

The use of RBF kernel decreased the Silhouette score of the clusters, but the homogeneity of the clusters and sub-clusters within improved. Figure 6 has already illustrated the distribution of *mouth* sentences using RBF kernel in the Spectral algorithm. It showed the separation between the metaphoric, metonymic and literal senses of *mouth* in the clusters and the closeness between face-related senses in clusters 2 and 4 and speaking-related senses in clusters 1 and 5.

Regardless of the parameters, the cohesion of the clusters increased after dimension reduction. Figure 10 summarizes the Silhouette scores of the k-means before and after using different 2D visualization methods for the four words examined in this study. It is evident that the cohesion of clusters considerably increased after the dimension reduction for all words. Also, the suggested best number of clusters differed across words and visualization methods. The highest Silhouette score was 0.838 for Spectral visualization of the sentences of *full*. For the same word, the Silhouette score for the MDS visualization was the lowest (0.392), although the two visualization methods recommended the same number of clusters. The visualization created by Spectral clustered the sentences closer to each other in two major classes. Most sentences following the pattern *full*+noun formed a cluster different from sentences following the pattern noun+V_{be}+*full of*+noun. Some sentences were sporadically distributed over the two clusters. However, they also showed some patterns, such as the collocations *full up* and *full to* and the pattern noun+V_{be}+*full*. Although the original senses of *full* in OD are 12, the Spectral visualization did not show sensitivity to the semantic differences between the sentences corresponding to the 12 senses. For instance, the metaphoric

senses of *full* (e.g., full of pain or joy) and the literal ones (e.g., full of books, clothes) are clustered in one category.

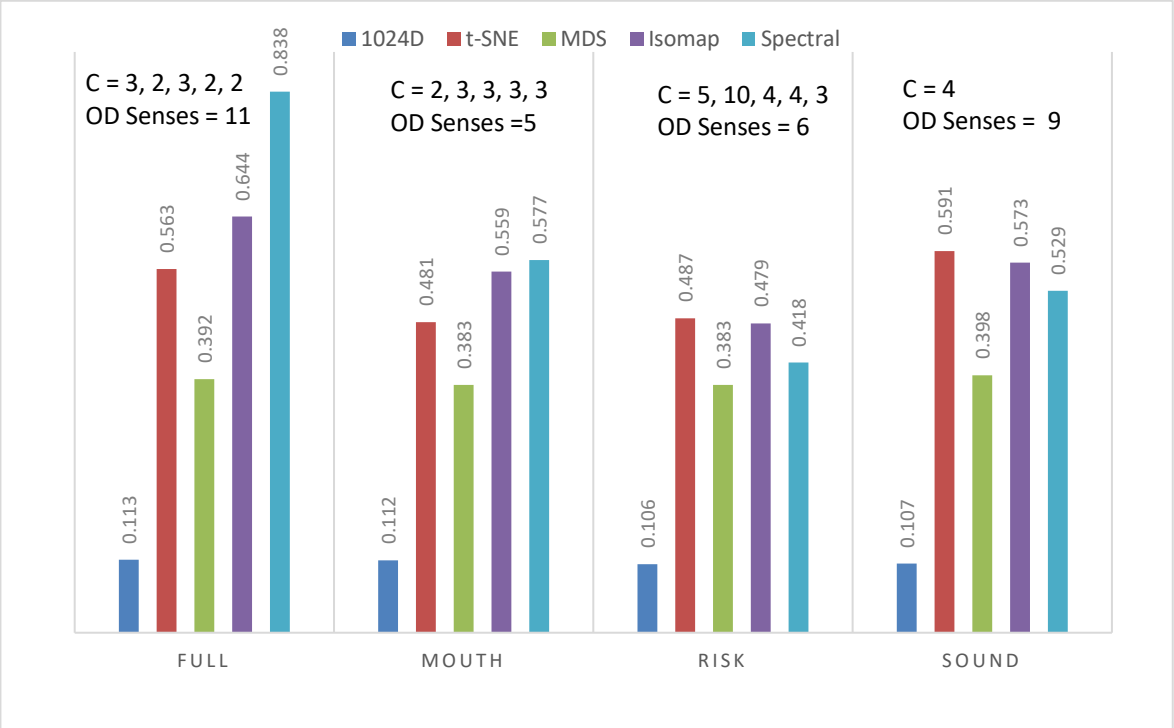

Figure 10: The highest Silhouette scores for the four studied words before and after dimension reduction

Several theoretical and computational approaches have been implemented in the literature to cluster dictionary senses into new categories. The clusters differed qualitatively and quantitatively according to the adopted approach. Whereas some studies depended on extensive qualitative analysis of dictionary data to improve the representation of senses for human users (e.g., Geeraerts, 2001; Lewandowska-Tomaszczyk, 2007; Molina, 2008), others aimed at improving the automatic performance of NLP tasks (for instance, Buitelaar, 1998, 2010; Ide & Wilks, 2007). Therefore, the number and members of the suggested clusters differed considerably.

Theory-based studies in lexicography highlighted the necessity of finding meaning relations among word senses (e.g., metaphoric and metonymic extensions of the literal senses), identifying the core literal meaning or meanings from which other meanings descend and organize word senses in homogenous categories that have always differed from those in the dictionaries. Although our study depended on distributional, rather than cognitive linguistic, approaches, the separation between the metaphoric, metonymic and literal senses of words such as *mouth* and *sound* was done automatically based on the distributional features of the word uses. Also, the uses of words with relevance to specific semantic fields (e.g., *risk* in financial domains, *mouth* to make

facial expressions, *full* with relevance to emotions) stood out in the automatically generated clusters.

The automatically generated clusters lumped several dictionary senses in the same cluster. It was most evident in the case of *full*, which had 11 fine-granular dictionary senses in our study. Yet, the different algorithms suggested 2 or 3 clusters only. Although the sub-clusters separated the metaphoric and the literal uses which were lumped in the dictionary, they also lumped the different levels of fullness which were split in the dictionary.

In almost all cases, the four algorithms reduced the number of OD's sense categories. Some dictionary distinctions were preserved within the sub-clusters (e.g., *sound* of music vs. *sound* of TV and radio), but others were lost (e.g. the four verbal senses of *risk*). Reducing the number of dictionary senses has been proposed in some NLP initiatives that prioritize the improvement of the quantitative indicators (the accuracy of word sense disambiguation). They, however, sometimes opt for solutions that are incompatible with the lexicographic practice, such as maintaining only meaning distinctions at the highest ontological levels, as discussed by Ide and Wilks (2007).

Our study aimed at combining extrinsic assessment of the clusters with qualitative analysis of their homogeneity so that the experiments can be relevant to both lexicographers and NLP scholars interested in sense-related tasks.

5. Conclusion

This study explored the possible use of 2D visualizations of contextualized word embeddings in lexicographic context, specifically sense delineation and example selection. It presented case studies for lexicographers to test the applicability of employing the suggested visualization methods in lexicographic investigations. Although the distributionally-created clusters did not correspond to the number of dictionary senses, they showed BERT's sensitivity to semantic and syntactic similarities between word uses.

Before dimension reduction, Silhouette scores of the *k*-means clusters were low, and so was the qualitative cohesion between the sentences in the cluster. Accordingly, providing lexicographers with distributionally-recommended clusters based on the original high-dimensional word embeddings are not helpful.

Visualizing BERT representations in 2-dimensional spaces using Spectral, t-SNE, Isomap and MDS algorithms showed quantitative and qualitative improvements that can be beneficial to lexicographers. For instance, not only the Silhouette scores of the *k*-means clusters increased, but also semantic and syntactic similarities appeared in the clusters and the manually identified sub-clusters within them.

Although the scope of the present study is limited to four words, to four dimension-

reduction methods and a single contextualized word embedding type (albeit a powerful one), we find these results novel and useful. The visualization of contextualized word embeddings of neologisms can help lexicographers identify their collocational patterns, POS usages and semantic preferences. Such patterns consistently appeared in the four case studies. Also, these visualizations can be helpful in enriching dictionary entries with additional, corpus-based examples; the closest BNC sentences to the OD examples mostly reflected very similar semantic and syntactic patterns in the four cases. In our charts, we also saw thematically-motivated clusters of BNC sentences that were ignored during exemplification of the OD headword (consider the uses of the word *mouth* in romantic literature), a situation which – when a representative corpus is used for the analysis – indicates a hiatus in the entry, which is not readily observable in concordances.

By taking advantage of the power of contextualized word embeddings and dimension reduction algorithms, we should be able to provide methods for lexicographers to explore and better understand the complex relationships between words and their meaning. These methods – enabled by current advances in Natural Language Processing – do not replace any subtask of the human "art and craft" of dictionary compilation, but they contribute to computer-assisted lexicography.

6. Acknowledgements

This publication was supported by the University of Debrecen Faculty of Humanities Scholarly Fund.

7. References

Abdelzaher, E. & Tóth, Á. (2020). Defining Crime: A multifaceted approach based on Lexicographic Relevance and Distributional Semantics. *Argumentum*, 16, pp. 44–63, https://doi.org/10.34103/ARGUMENTUM/2020/4.

Buitelaar, P. (1998). *CORELEX: Systematic polysemy and underspecification.* PhD thesis. Waltham, Massachusetts: Brandeis University.

Buitelaar, P. (2010). Ontology-based semantic lexicons: Mapping between terms and object descriptions. In C. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari & L. Prevot (eds.) *Ontology and the lexicon: A natural language processing perspectiv*e. Cambridge: Cambridge University Press, pp. 212–223.

Demsar J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., Mozina, M., Polajnar, M., Toplak, M., Staric, A., Stajdohar, M., Umek, L., Zagar, L., Zbontar, J., Zitnik, M., Zupan, B. (2013.) Orange: Data Mining Toolbox in Python. *Journal of Machine Learning Research*, 14, pp. 2349−2353.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Fillmore, C. & Atkins, S. (1992). Towards a frame-based organization of the lexicon: the semantics of RISK and its neighbors. In A. Lehrer and E. Kittay (eds.) *Frames, fields, and contrasts: New essays in semantic and lexical organization*, pp. 75–102.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In J. R. Firth (ed.): *Studies in linguistic analysis*. Oxford: Basil Blackwell, pp. 1–32.

Geeraerts, D. (2001). The definitional practice of dictionaries and the cognitive semantic conception of polysemy. *Lexicographica*, 17, pp. 6–21.

Groenen, P. J. F. & Borg, I. (2014). The Past, Present, and Future of Multidimensional Scaling. In J. Blasius & M. Greenacre (eds.) *Computer Science and Data Analysis Series: Visualization and Verbalization of Data*. CRC Press: Boca Raton, FL, USA; Taylor & Francis Group: Didcot Melton Park/London/Abingdon, UK, pp. 95–117.

Hornby, A. S. (1948). *Idiomatic and Syntactic English Dictionary. Institute for Research in Language Teaching.* Tokyo: Kaitakusha.

Ide, N. & Wilks, Y. (2007). Making sense about sense. In E. Agirre & P. Edmonds (eds.) *Word sense disambiguation*. Dordrecht: Springer, pp. 47–73.

Jatowta, A., Tahmasebib, N. & Borinb, L. (2021). Computational approaches to lexical semantic change: Visualization systems and novel applications. *Computational approaches to semantic change*, 6(311).

Kilgarriff, A. (2007). Googleology is Bad Science. *Computational Linguistics*, 33(1), pp. 147–151, http://doi.org/10.1162/coli.2007.33.1.147.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), pp. 7–36, http://doi.org/10.1007/s40607-014-0009-9.

Lakoff, G. (1987). *Women, fire, and dangerous things*. Chicago: University of Chicago Press.

Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). *Language research*, 28(1), pp. 1–13.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1), pp. 1–31.

Lewandowska-Tomaszczyk, B. (2007). Polysemy, prototypes, and radial categories. D. Geeraerts & H. Cuyckens (eds.) *The Oxford handbook of cognitive linguistics*. Oxford: Oxford University Press, pp. 139–169.

Molina, C. (2008). Historical dictionary definitions revisited from a prototype theoretical standpoint. *Annual Review of Cognitive Linguistics* 6(1), pp. 1–22.

Montes, M. & Heylen, K. (2022). Visualizing distributional semantics. In D. Tay & M. X. Pan (eds.) *Data Analytics in Cognitive Linguistics: Methods and Insights*. Berlin, Boston: De Gruyter Mouton, pp. 103–137.

Ng, A. Y., Jordan, M. I. & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems (NIPS),* 14, pp. 849–856.

OD: Oxford Learner's Dictionaries, online version, accessed at:

https://www.oxfordlearnersdictionaries.com (7 February 2023).

Ostrovsky, R., Rabani, Y., Schulman, L. J. & Swamy, C. (2006). The Effectiveness of Lloyd-Type Methods for the k-Means Problem. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06). IEEE.,* pp. 165–174.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume 1*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227-2237.

Radford, A, Narasimhan, K, Salimans, T. & Sutskever, I. (2018). Improving language understanding by generative pre-training. Available at https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

Rychlý, P. & Kilgarriff, A. (2007). An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In S. Ananiadou (ed.) *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions.* Stroudsburg, PA: Association for Computational Linguistics, pp. 41–44, http://doi.org/10.3115/1557769.1557783.

Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science,* 290(5500), pp. 2319–2323.

van der Maaten, L. & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 1, pp. 1–48.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. & Polosukhin, I. (2017). Attention is all you need. In: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett (eds.) *Advances in Neural Information Processing Systems*, 30, pp. 5998–6008.

Wiedemann, G., Remus, S., Chawla, A. & Biemann, C. (2019). Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. In *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019.*