Unveiling Word Meaning: A statistical study on primitive semantic features

Davide Testa[1,2],

[1] Fondazione Bruno Kessler (FBK), Trento
[2] University of Rome La Sapienza

Abstract :

Semantic representation approaches have long struggled with defining primitive features that adequately capture the essential components of meaning. Traditional methods often face challenges due to the complexity of these features. The brain-based componential approach, which links semantic features to specific neural systems, offers a more biologically plausible and mechanistic account of conceptual representation. This study aims to validate the robustness and validity of this approach by addressing two research questions: whether similarity structures can be identified through semantic features and whether morphosyntactic classes can be distinguished based solely on these features. Using the dataset from Binder et al. (2016), an extensive statistical analyses has been conducted. Both unsupervised and supervised learning analyses confirmed the effectiveness and well-structured nature of these semantic features. Notably, the unsupervised analysis revealed a new and interesting observation: the significant role of two diametrically opposed semantic dimensions, Concreteness and Abstractness, in the construction and representation of meaning. The findings demonstrate the potential of the brain-based componential approach in representing and distinguishing semantic categories, also suggesting its utility for developing advanced AI systems capable of human-like language processing.

1 Introduction

Semantic representation theory investigates how meanings and concepts are encoded in the mind and brain. A key approach within this field is the componential approach, which posits that concepts are composed of sets of primitive features or attributes. These features aim to capture the essential components of meaning, but traditional theories often struggle with defining truly primitive features, as these components can themselves be complex. A more recent development is the brain-based componential approach, which links semantic features to specific neural systems involved in perception, action, and other modal experiences. This embodied perspective suggests that conceptual knowledge is grounded in the sensory and motor systems through which it is acquired (Binder et al., 2016). The advantages of using a brain-based componential approach are many. By grounding semantic features in neurobiological processes, this method provides a more biologically plausible and mechanistic account of how concepts are represented and learned. This approach also allows for a more integrated understanding of how different types of experiences (sensory, motor, affective) contribute to concept formation. Additionally, leveraging this approach for machine representation of meaning holds significant potential. It can enhance the development of more sophisticated and human-like artificial intelligence systems capable of understanding and processing language in a way that mirrors human cognitive processes. Such systems could improve natural language processing, machine translation, and even human-computer interaction by providing machines with a deeper, more nuanced understanding of meaning.
Taking in mind this theoretical foundation, the current work seeks to validate the robustness and validity of the brain-based componential approach proposed in Binder et al. (2016). To achieve this, the study addresses two main research questions:

1. Can similarity structures be identified by the semantic information conveyed by semantic features?
2. Can morpho-syntactic classes be distinguished solely based on semantic information?

By answering these questions, this work aims to confirm the validity of the brain-based componential approach and demonstrate its power in representing concepts and conveying crucial information for language understanding.
Concretely, the work presented in the following sections consists of a series of statistical analyses conducted using the dataset built by Binder et al. (2016) as a reference point. This dataset provided a robust foundation for examining the salience and structure of various experiential attributes, allowing for a detailed investigation into the potential of the brain-based componential approach to accurately represent and distinguish semantic categories.
The paper is organized as follows. Section 2 discusses previous works in this specific research area. Section 3 presents the design and structure of the dataset used for the analysis. In Section 4, the methods and the approaches used for the experiments are discussed. Section 5 reports and discusses the results, while Section 6 and 7 show where these tests led and how these results may pave the way for further research.

2 Related Works

Classical category theory defines concepts by binary features necessary and sufficient for category identification. For example, the concept of a bird includes features such as wings, feathers, beak, and flying. However, it has been recognized that most categories have fuzzy boundaries due to varying degrees of prototypicality (Rosch et al., 1976). Extensive research using feature generation tasks has documented the features of entities and their importance for a given concept, helping in assessing similarity between concepts and grouping them into hierarchical categories (McRae et al., 1997; Cree et al., 2003).
A limitation of standard approaches is that the features used to define conceptual content are often complex concepts themselves. Physical features like wings and feathers, while components of larger entities, are not primitive since they require perceptual and verbal experience to be learned. This issues poses a significant challenge for feature-based theories, as the set of possible features is vast for even known objects and actions (e.g., flying, swim-

ming, running). One significant limitation is the lack of a known relationship between verbal features and neurobiological mechanisms. There are no neural systems specifically dedicated to representing complex features like feathers or wings. According to Bowers (2009), even if neurons dedicated to specific concepts exist, this does not explain how stimuli lead to their activation. Understanding why concepts are organized in the brain as they are and how they are learned is crucial. Abstract "concept cells" fail to address how concepts are grounded in sensory-motor experience, which is crucial for referencing the external environment (Harnad, 1990). Thus, traditional feature-based semantic theories aim to describe entities, their similarity structures, and category groupings but do not address how this information is organized in the brain. In contrast, embodiment theories of knowledge representation analyze conceptual content in terms of sensory, motor, affective, and other experiential phenomena, and their corresponding neural representations. Jackendoff (1992) started, within its *Conceptual Semantics* theory, emphasizing experiential primitives like space, time, and causality. Subsequent work has expanded on these ideas to include a wider range of sensory, motor, and affective attributes and their neural correlates. Studies on modality-specific contributions to conceptual knowledge have often focused on concrete concepts. Lynott et al. (2013) gathered ratings on sensory associations for concrete adjectives and nouns, while Gainotti et al. (2013) added dimensions like shape, color, and manipulation experience. Hoffman et al. (2013) then contributed to share the importance of using modality-specific attribute ratings for lexical-semantic processing. Instead, more recent studies have highlighted the role of affective and social experiences in abstract concept acquisition and embodiment (Katja Wiemer-Hastings et al., 2005; Borghi et al., 2011). Finally, Troche et al. (2014) provided further insights into how abstract nouns are rated on dimensions such as emotion, polarity, social interaction, morality, and space. The advent of brain imaging tools has increased interest in understanding how concepts are represented in human neural systems. Evidence supports the idea that these representations are at least partly embodied in the perception, action, and other modal neural systems through which concepts are acquired (Meteyard et al., 2012).

Starting from this, the aim of Binder et al. (2016) was to develop a comprehensive conceptual representation based on known modalities of neural information processing. This representation captures aspects of experience central to the acquisition of both object and event concepts, as well as abstract and concrete concepts, allowing to represent the meaning of a set of lexical items through this approach.

3 Dataset

The dataset was taken from Binder et al. (2016) and the theory underlying its construction starts from the assumption that our knowledge is inherently multimodal and it comes from a multimodal experience. For these reasons, its main goal represents an attempt to describe word meaning by the means of a set of semantic multimodal information.

Units. Datapoints consists in lexical items (i.e., Words). In the specific case of the dataset, Binder et al. (2016) wanted to analyze the semantic behaviour of Nouns, Verb and Adjectives.

Features. Based on this componential semantic approach, features represent the ratings of the relevance of each semantic component to word meaning. Basically this variables can be seen as primitive features for the analysis of conceptual content and thus, their sum should create the meaning of a single lexical units. Concretely, they are experiential attributes coming from some *Embodiment theories of knowledge representation*[1] (Louwerse et al., 2005) and which provide a set of primitive features for the analysis of such conceptual content, allowing the exploration of the (mental) dimension of experience. The ratings used in this dataset and for the next statistical studies were obtained by sophisticated brain-imaging tool such as fMRI and correspond to a selected group of primitive features that are shown in Table 1 and 2 with reference to their semantic domain.

Moreover, such a dataset contains also other types of qualitative features, coming from previous clustering analysis or manually labelling tasks.

Table 1
Sensory and motor semantic features organized by domain

Domain	Feature
Vision	Vision
Vision	Bright
Vision	Dark
Vision	Colour
Vision	Pattern
Vision	Large
Vision	Small
Vision	Motion
Vision	Biomotion
Vision	Fast
Vision	Slow
Vision	Shape
Vision	Complexity
Vision	Face
Vision	Body
Somatic	Touch
Somatic	Temperature
Somatic	Texture
Somatic	Weight
Somatic	Pain
Audition	Audition
Audition	Loud
Audition	Low
Audition	High
Audition	Sound
Audition	Music
Audition	Speech
Gustation	Taste
Olfaction	Smell
Motor	Head
Motor	Upper Limb
Motor	Lower Limb
Motor	Practice

[1] Many of them have already been presented above in Section 2

Table 2
Spatial, Temporal , Causal, Social, Emotion, Drive and Attention features

Domain	Feature
Spatial	Landmark
Spatial	Path
Spatial	Scene
Spatial	Near
Spatial	Toward
Spatial	Away
Spatial	Number
Temporal	Time
Temporal	Duration
Temporal	Long
Temporal	Short
Causal	Caused
Causal	Consequential
Social	Social
Social	Human
Social	Communication
Social	Self
Cognition	Cognition
Emotion	Benefit
Emotion	Harm
Emotion	Pleasant
Emotion	Unpleasant
Emotion	Happy
Emotion	Sad
Emotion	Angry
Emotion	Disgusted
Emotion	Fearful
Emotion	Surprised
Drive	Drive
Drive	Needs
Attention	Attention
Attention	Arousal

3.1 Dataset structure

Since the dataset has been build and used previously for similar studies in Binder et al. (2016), it was originally a little bit more complex, having some more features than those needed for the scopes of this research[2]. For this reason, the original number of features have been reduced and, after removing some duplicates inside the units, the final version of the dataset is composed of 534 units, representing lexical items (i.e., Nouns, Verbs and Adjectives) and 66 features: 65 primitive semantic features[3] plus a qualitative feature (i.e., *Type*) representing the lexical type of the unit from a semantic perspective. This last one allows us to understand more quickly the data items composing the dataset in terms of morpho-syntactic classes as shown in Table 3[4].

Table 3
Dataset composition in terms of lexical type

Type	Morpho-syntactic class	**Number of items**
Things	Nouns	434
Actions	Verbs	56
Properties	Adjectivs	44

4 Methods

Understanding the intricate nuances of word meaning and lexical semantics involves a multifaceted approach. For this reason, this section outlines the methodological framework employed in this study, which aimed to explore the semantic landscape of the meaning underlying the given lexical items while trying to test simultaneously the goodness of such a componential approach.
What has been done is divided into three distinct subsections: *Data pre-processing*, where the steps taken to pre-process and structure the dataset for analysis plus some preliminary study are detailed; *Unsupervised Learning Analysis*, where unsupervised learning techniques are used to uncover latent semantic patterns and groupings within the dataset; and *Supervised Learning Analysis*, wherein supervised learning algorithms are employed to predict and classify lexical types based on the primitive semantic features.

4.1 Data pre-processing

The quality and integrity of the dataset are ensured by data pre-processing, a crucial step in any data analysis pipeline. Thus, this section discusses the key pre-processing steps undertaken in this study in order to make the dataset more suitable for the next analysis.
First, the dataset was examined for the presence of NaNs using functions that identify and count these missing values across dif-

[2] Some of these features were not strictly related to the componential semantic approach. Binder et al. (2016) conducted other types of statistical analysis and such features were the quantitative result of some artificial-metric
[3] The ones previously presented in Table 1 and 2
[4] Some of the adjectives were originally referred as *Type: state*. Since this can be read as a specification, the label for that adjectives has been brought back to the more general *properties*. This last update was done also in order to make the computation more efficient for the next analysis.

ferent features. This assessment revealed several features with missing values that could potentially skew the results and reduce the robustness of subsequent analyses. To mitigate the impact of these missing values, a mean strategy was implemented. This involved replacing each NaN with the mean value of the respective feature, thus, maintaining the overall distribution and central tendency of the data, and minimizing the introduction of bias. Moreover, this step ensured that the analyses were not adversely affected by missing data, thereby enhancing the validity of the findings related to the semantic similarity structures within the features.

Then, during the preliminary data analysis, a check was conducted to assess the skewness of the data. This examination revealed that the dataset exhibited a pronounced right-skewed distribution, indicating the presence of many outliers. Such skewness can significantly impact subsequent analyses, potentially leading to biased results and misinterpretations. To mitigate this issue and ensure the robustness of the analytical methods employed, a log-scale transformation was applied to the data. This transformation helped to normalize the distribution, reduce the influence of outliers, and enhance the reliability of the subsequent analyses. Figure 1 shows an example of the different data distribution of a single features (i.e., *Color*) before and after log-transformation, respectively on the upperside and lowerside of them. After this changes, the skewness means changed from 1.34 to 0.59.

The final step was the normalization of data which is another critical aspect of data pre-processing since the use of normalized data is essential for many analytical methods. By standardizing the features to have a mean of zero and a standard deviation of one, normalization ensures that each feature contributes equally to the analysis, preventing features with larger ranges from dominating the results.

Once having done this pre-processing step and before starting with the actual analyses introduced in 4.2 and 4.3, it was useful to conduct a preliminary study with respect to the available data. The purpose is intended to see how the data are distributed and the possible presence of relationships among them. Therefore, a check was made to understand the possible presence of multicollinearity relationships among features through the use of a correlation matrix. Results in Figure 2 shows that it is possible to identify a linear dependency among many of our predictors. Such behavior makes it possible to highlight macro-dimensions of correlations that identify groups of features well defined by specific semantic domains. The first major group of highly correlated features is related to the somatosensory domain, which includes mainly vision and somatic features. After that, we can also find quite clearly the group of features related to the auditory domain. On the other hand, the correlation identified between temporal and causal features (i.e., Temporal and Causal domain) is interesting. Finally, there is a subset of features belonging to the emotion domain that turn out to be well correlated.

Such behavior described in the correlation matrix certainly has consequences toward subsequent statistical analyses. The choice, however, was not to adopt any resolution strategy (e.g., to merge the highly correlated features into a unified macro-feature) since the purpose of such a study is precisely to see to what extent the intervention of the single predictor is valid in defining the meaning of the single word.

4.2 Unsupervised Learning analysis

Since unsupervised learning methods are used to uncover hidden patterns, groupings, or structures within the data, this approach is usually particularly useful for exploratory data analysis, anomaly detection, and dimensionality reduction. In this study, such methods were employed to address a specific research question: "*Can we identify similarity structures conveyed by the semantic information of the features?*". To explore this, clustering techniques such as k-means, hierarchical clustering and biclustering were initially applied[5]. These methods aimed to group the data based on inherent similarities, potentially revealing underlying structures within the features and confirming the goodness of such a (semantic) componential approach. In addition, biclustering simultaneously clusters rows and columns of a matrix, identifying submatrices where rows exhibit similar behavior across specific columns. This approach is particularly advantageous in high-dimensional data analysis, as it can uncover local patterns and relationships that traditional clustering methods might miss, providing deeper insights into complex datasets while computing also a feature clustering. After that, based on clustering results, a Principal Component Analysis (PCA) has been used as an attempt to structure a denoising approach by reducing the dimensionality of the data while retaining the most significant variance, thereby enhancing the clarity of the subsequent clustering analysis. Thus, it was concretely used for deriving a low-dimensional set of features from a large set of variables, where each dimension found by PCA is a linear combination of the p features.

4.3 Supervised Learning analysis

Differently from the previous section, the supervised learning component of this project aimed to address the following research question: "*Can morphosyntactic classes be distinguished solely based on semantic information?*". To investigate this, a classification task was performed using multinomial logistic regression with a Lasso penalty. The objective was to classify data items into one of three labels in the feature "Type": action, property, or thing, which correspond to the lexical classes of verbs, adjectives, and nouns, respectively. The model was trained on the dataset to predict the appropriate lexical class for each data item based on its semantic features. By focusing on semantic information, the analysis aimed to determine the extent to which these features could reliably indicate the morphosyntactic category of a word.

The R *glmnet* package (Friedman et al., 2010) was employed for this task, taking advantage of its capability to handle multinomial regression with regularization. Specifically, Lasso is a regularization technique (also knows as L1 regularization) that promotes sparsity within the model while trying to avoid issues of multicollinearity and overfitting within datasets. Thus, it enhances the accuracy of statistical models and its interpretability by selecting only the most relevant features.

Additionally, a feature selection task was conducted as a result of applying the Lasso penalty in the regression. This involved examining the most important features selected by the model. Identifying these key features is a good way to provide deeper insights into which semantic properties were most influential in distinguishing between the morphosyntactic classes, thereby enhancing the un-

[5] Section 5 shows the results and the configuration of the best method among all those tried.

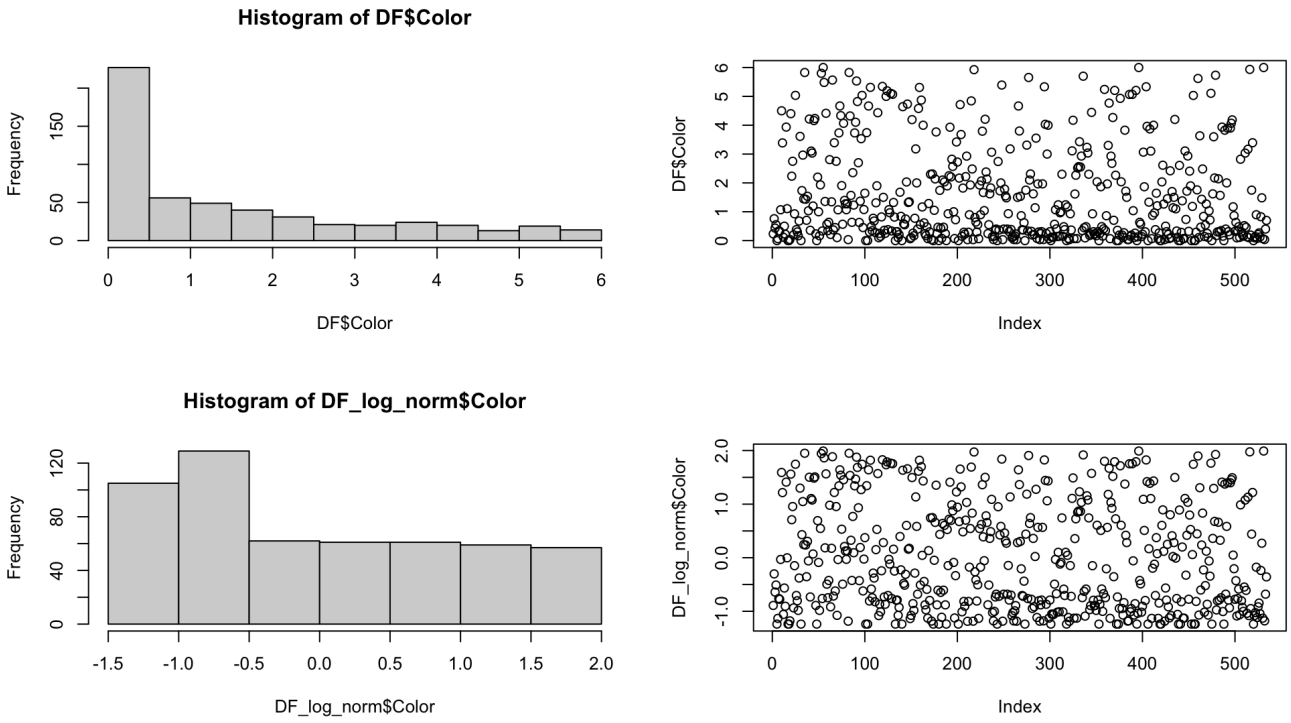


Figure 1. Data Distribution of *Color* before and after log-transformation

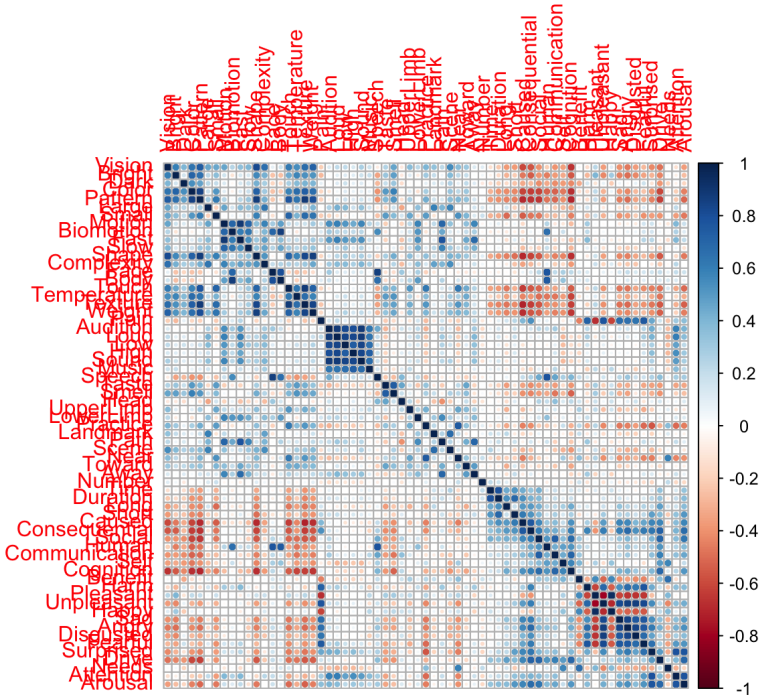


Figure 2. Correlation matrix of the 65 features

derstanding of the relationship between semantic information and lexical categories.

5 Results

This section presents the findings from the various analyses conducted to address the research questions of this study. Each subsection provides the results for the specific methods applied, highlighting the key outcomes and insights gained from the analysis.

5.1 Clustering

In order to uncover underlying patterns and similarity structures within the data several clustering techniques were employed, including k-means, hierarchical clustering, and biclustering. Results allowed to explore how the data items grouped together and whether these groupings aligned with the expected semantic relationships. Each clustering method provided distinct perspectives on the structure of the data, helping to identify sometimes meaningful clusters and assess the effectiveness of each technique. The clustering attempts presented turn out to be very different from each other, since they are based on different approaches, but with regard to k-means and hierarchical clustering the steps adopted were consistent and systematic:

- Determine the **Optimal Number of Clusters**: Techniques such as the elbow method, silhouette method and Hartigan Index were used to identify the optimal number of clusters based on statistical criteria.
- Apply **Clustering Algorithm**: Once the optimal number of clusters was determined, the clustering algorithms were applied to partition the dataset into those clusters.

By following this approach, the clustering analysis for these two methods was guided by statistical principles, leading to more meaningful and interpretable results.
The following paragraphs detail the outcomes of each clustering approach by presenting the various methods used with the different configurations that led to the best results, although many more attempts were made than those in this paper.

K-means Initial efforts to cluster the dataset into 2 groups (Figure 3) did not yield informative results, even though the methods for finding the optimal number of clusters all agreed with a value of k = 2, as shown in Figure 4. While this solution could not appear as the best one, Figure 4(c) shows the most similar behavior in terms of silhouette values for k = 10 (0.19 for k= 2 against 0.12 for k = 10). However, despite subsequent exploration with k = 10 (Figure 5), the clusters obtained were still somewhat chaotic and lacked clear interpretability, although a semantic-based partitioning of the data is beginning to be observed.

Hierarchical Clustering Since k-means clustering was not efficient in producing meaningful partitions of the data, hierarchical clustering was employed as an alternative method. Similar to the k-means approach, the optimal number of clusters was determined using three different methods, which again confirmed the k = 2 as the best one, but suggested k = 10 as a viable options based on the average silhouette scores. The dendrogram obtained through this agglomerative hierarchical clustering by using a complete linkage method[6] was cut using a *cuttree()* command with k = 2 and k = 10 to evaluate and compare the clustering performances. While the partitioning with k = 2 did not produce consistent results, clustering with k = 10 (Figure 7) significantly improved the identification of semantically well-based groups. Although not perfectly, the k = 10 clusters were much better at capturing the underlying semantic structures within the data, not only with reference to the two clusters solution found with hierarchical clustering, but also with the corrispondent clustering with k-means. Specifically, here with k = 10, some rough semantic classes began to emerge, which could be with good confidence linked to notions such as Food, Animals, Natural Events, and others.

Biclustering After trying the previous two clustering approaches, the best solution, using the source data without further handling, turned out to be biclustering. The results presented here correspond to the outputs obtained with the R *isa-2 package* (Csardi, 2023), which have been the best performing in this task compared to other libraries tested[7]. The algorithm used was able to find 199 clusters that were consistent from a semantic point of view, both in terms of the lexical items in the dataset and the features that compose them. Figure 8 represents the top 8 clusters identified, which, in order of presentation from left to right, can be ranked on a semantic basis as shown in Table 4. Theoretically, the use of biclustering should allow for a unified look at the entire dataset, divided into clusters that have correlated rows and columns. However, in this case, this was not fully achievable due to the presence of significant instances of overlapping. This overlap indicates that some lexical items recur in different clusters, reflecting the complexity and multifaceted nature of the dataset. Nonetheless, biclustering provided the most semantically coherent grouping among the methods tested so far.

Table 4
Semantic classification of the best 8 biclusters identified

Bicluster	Semantic domain
Bicluster1	Time
Bicluster34	Food and Drinks
Bicluster43	Animals
Bicluster49	Sounds
Bicluster57	Locations
Bicluster76	Negative Events and Feelings
Bicluster20	Human beings (Job)
Bicluster79	Practical Tools

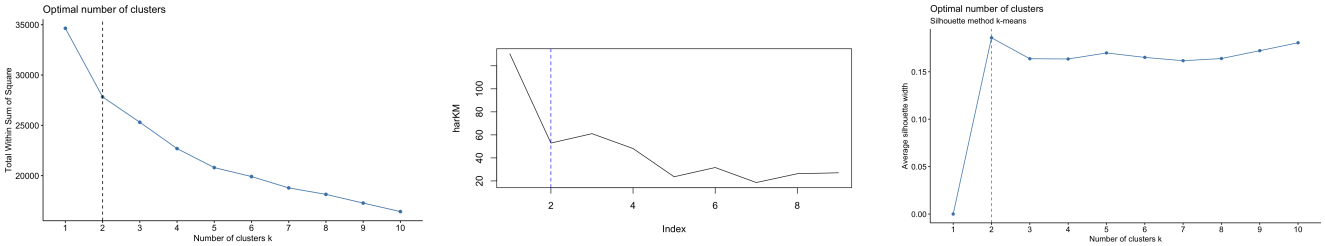
5.2 PCA and clustering

Principal Component Analysis (PCA) was employed as another attempt to obtain more interpretable and well-defined results in the clustering analysis. Theoretically, by reducing the dataset to its most significant components, PCA helps to eliminate noise and redundancy, thereby facilitating the identification of underlying pat-

[6] Other linkage methods were tried but they did not lead to consistent results

[7] Among the several packages in R that allows to do biclustering, the Iterative Signature algorithm used by *isa2-package* turned out to be the best one both for the clustering retrieval and the visual presentation of data.

Figure 3. Kmeans with K=2



((a)) Within cluster dissimilarity/distance **((b))** Hartigan Index **((c))** Average Silhouette

Figure 4. Optimal number of clusters: K-means

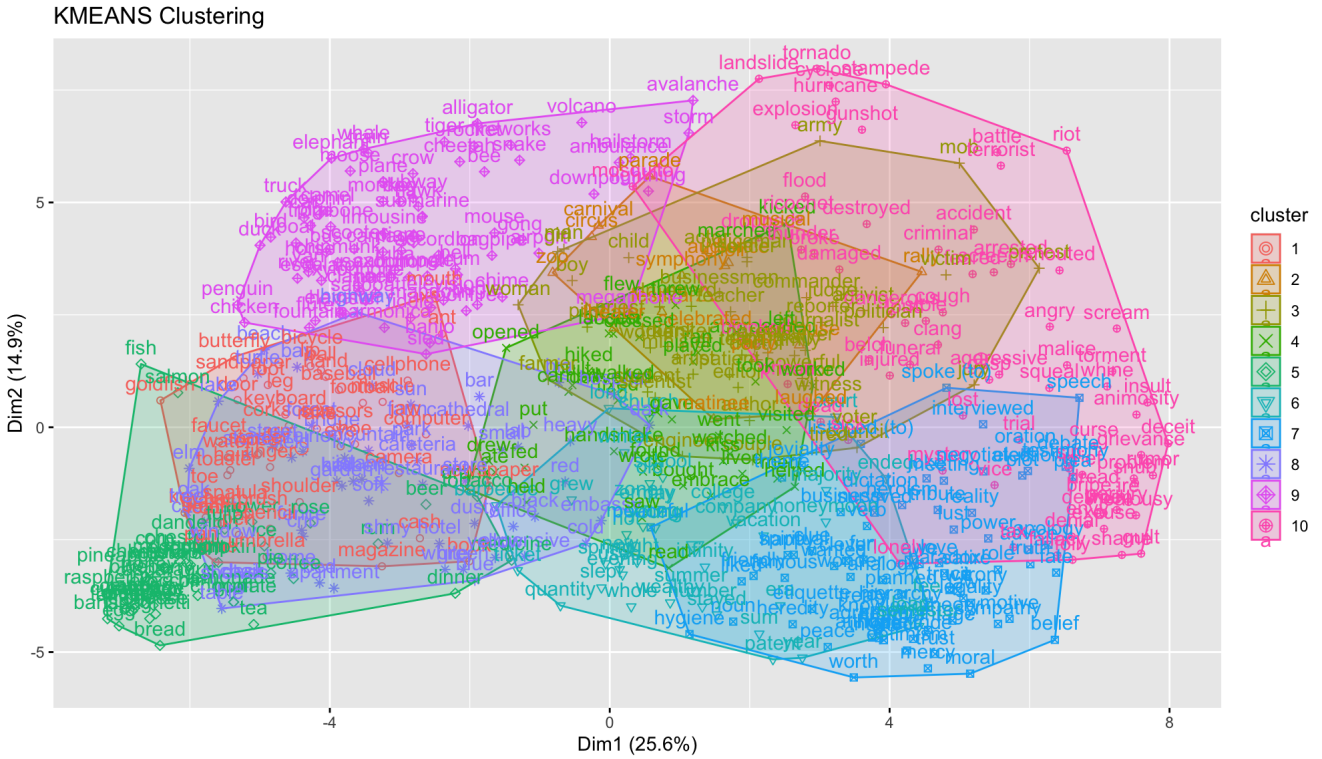
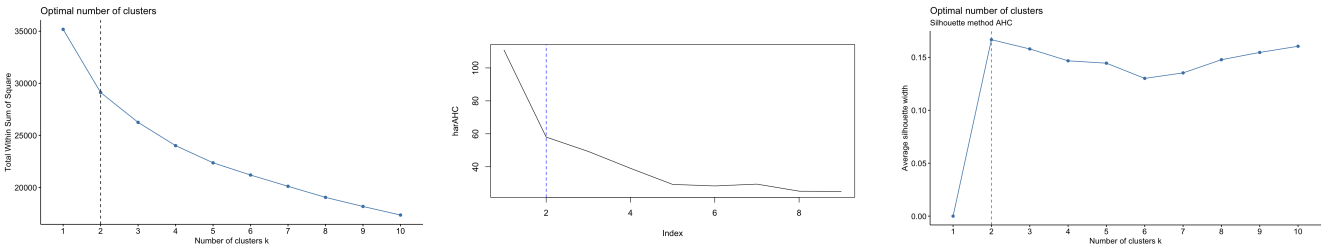


Figure 5. Kmeans with K= 10



((a)) Within cluster dissimilarity/distance **((b))** Hartigan Index **((c))** Average Silhouette

Figure 6. Optimal number of clusters: Hierarchical Clustering

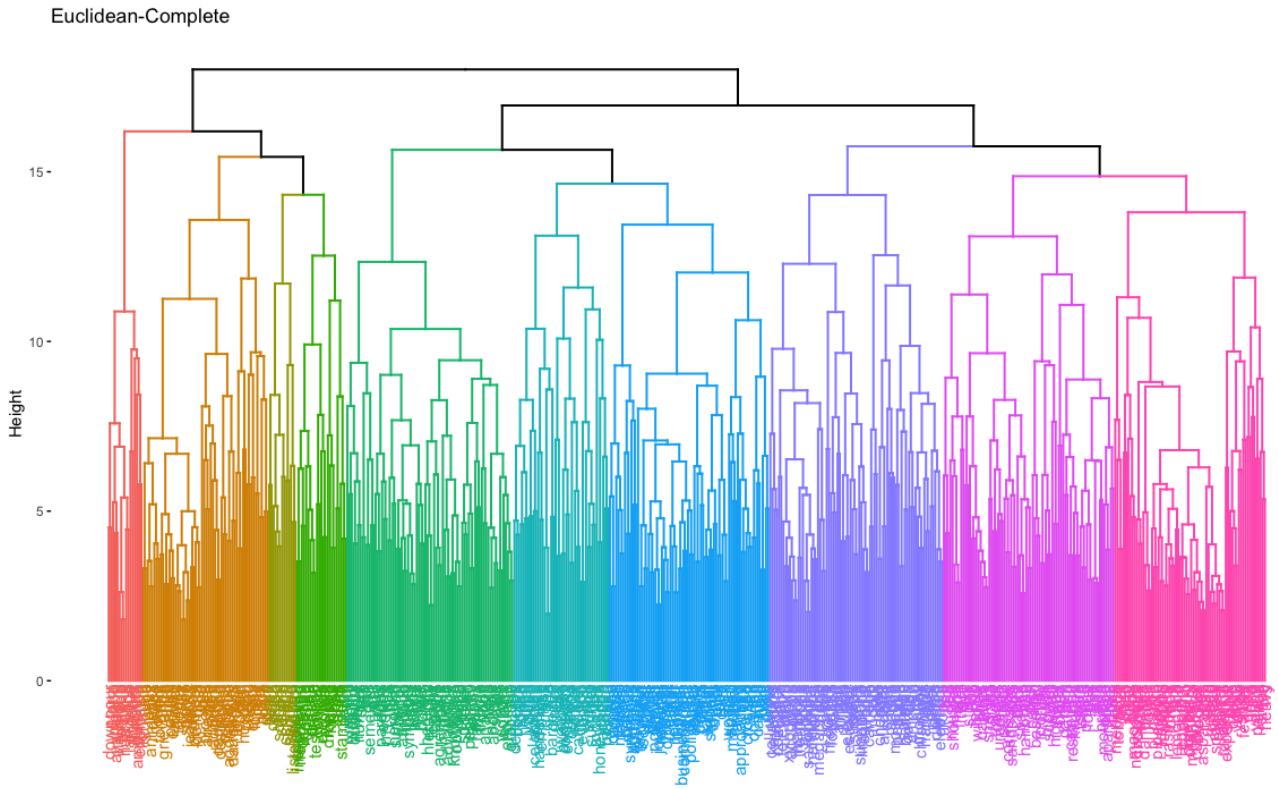


Figure 7. Agglomerative Hierarchical Clustering with K = 10

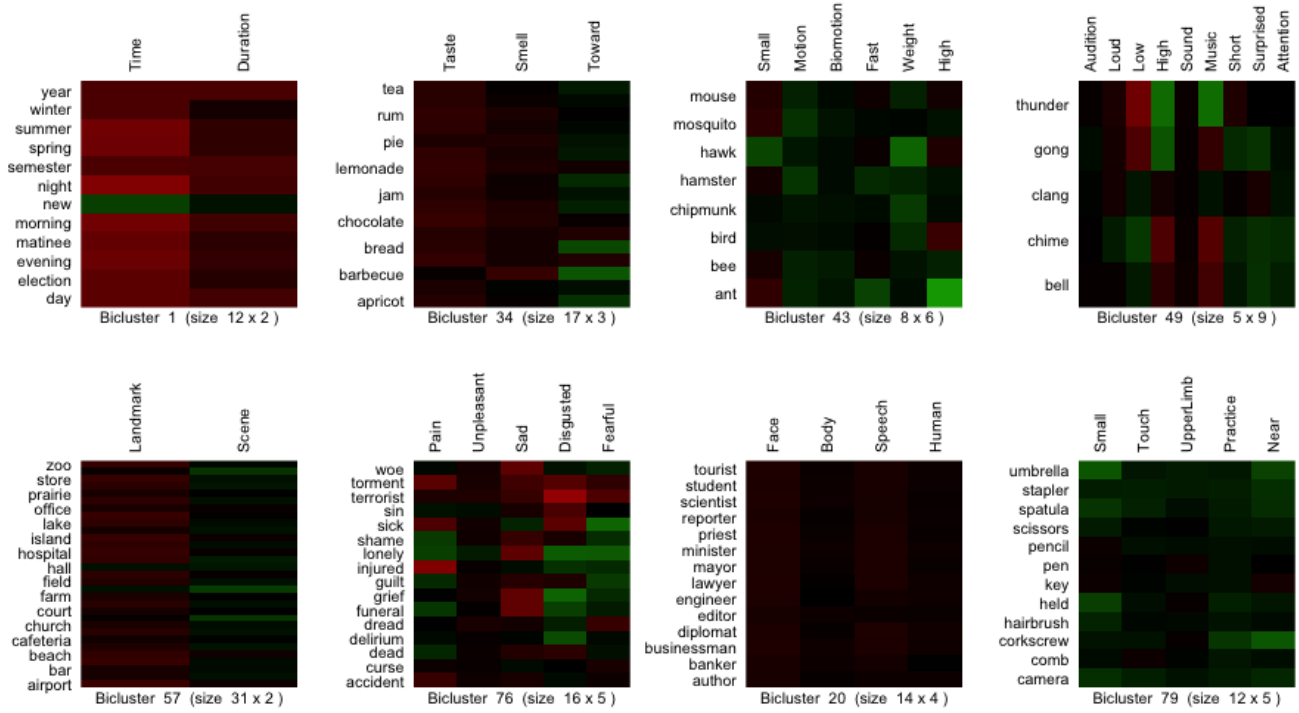


Figure 8. Eight most semantically consistent biclusters identified by the isa-2 algorithm

terns. After applying PCA, the percentage of variance explained by the individual components was examined. Figure 9(a) and 9(b) demonstrate that 80% of the data variability is captured by the first ten dimensions alone, with the first three being the most representative.

In the analysis of the loadings of the principal components (i.e., the eigenvectors representing the direction of the PCs), the first two dimensions were projected onto a correlation circle or variable plot (Figure 10). This revealed and interesting thing: the most important features in the first and second dimensions can be approximated to two quite well-distinguished and diametrically opposed semantic dimensions, since all the *Temporal*, *Causal*, *Social*, *Emotion* and *Attention* semantic components contribute in the definition of the first principal component, while the *sensory-motor* and *spatial* features in the definition of the second one. This observation was then further confirmed by Figure 11, which highlights the most significant features conveyed by dimensions 1 and 2.

Given these findings and that these first two principal components are always the most informative one, the two clustering methods (i.e., k-means and hierarchical clustering) were subsequently applied to these dimensions. Although both methods were tried, hierarchical clustering with the complete linkage method proved to be more accurate. Assuming k = 2, the resulting clusters represented abstract versus concrete lexical items with good approximation. When attempting to replicate the earlier clustering steps with k = 10, even more accurate clusters emerged in terms of semantic domains. Further testing revealed that when starting with k = 10 and increasing to k = 28 (i.e., the number of semantic classes identified in Binder et al. (2016)), the accuracy and precision in identifying semantic macro-categories improved. However, this increase came at the expense of the average silhouette score, which decreased from 0.42 (for k = 2) to 0.37 (for k = 10) and 0.35 (for k = 28). Considering the better accuracy from a qualitative perspective and the marginal difference in silhouette scores, the optimal balance was found to lie between k = 10 and k = 28. This balance achieved a more refined clustering that accurately captured the semantic macro-categories while maintaining a reasonable silhouette score.

5.3 Multinomial Regression and Feature Selection

This section presents the results of employing a multinomial regression approach to classify the morphosyntactic class of each lexical item within the dataset. Various configurations were explored before arriving at the most effective approach for classification. Initially, a multinomial regression model was trained using all 65 features of the dataset in their original structure. Cross-validation was employed to search for the optimal lambda value, which was then used to train the final model[8]. The same procedure was repeated using the results of PCA, considering both the first 2 dimensions and 10 dimensions.

Despite achieving high accuracy values (0.94 for the configuration with 65 features, 0.81 using 2 principal components, and 0.86 using 10 principal components), the confusion matrices revealed a consistent issue. The models tended to classify the category "thing" well but struggled with other categories. Such a problem is due to the dataset's imbalance in terms of the label *Type* which has been used for the prediction, as previously illustrated in Table 3. Consequently, all three models exhibited overfitting behavior, confirmed by studying the error rate evolution with a learning curve.

To address the imbalance issue, an under-sampling approach was attempted. The items corresponding to the label *Thing* were randomly reduced from 434 to 60, creating a new balanced subdataset (Table 5).

Table 5
Subdataset composition in terms of lexical type after applying the balancing approach

Type	Morpho-syntactic class	**Number of items**
Things	Nouns	60
Actions	Verbs	56
Properties	Adjectivs	44

This balanced dataset was then used to reproduce the classification task with the same steps as previous attempts. This approach proved to be the most effective, producing a classification accuracy of 0.94. Moreover, the resulting confusion matrix demonstrated a well-structured classification (Table 6) and indicated an improved, though not perfect, reduction in overfitting behavior as observed through the learning rate.

Table 6
Confusion Matrix based on the classification of the new balanced dataset

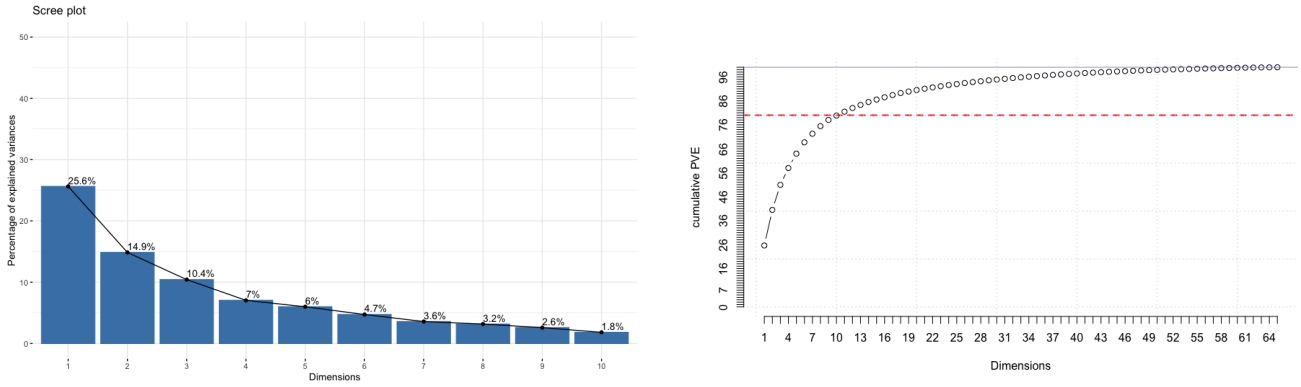
Prediction	*Thing*	*Action*	*Property*
Thing	18	16	13
Action	0	14	1
Property	0	2	12

Finally, as a consequence of applying a Lasso penalty, the non-zero output coefficients of the regression were used for feature selection. Table 4 displays the most important features selected and used for the classification task. Only 24 out of the initial 65 features were found to be meaningful and useful for classifying the lexical items (Table 7).

6 Discussion

The statistical analyses conducted in this study effectively addressed the initial research questions. In the unsupervised learning part, the k-means clustering algorithm, despite its common usage for partitioning datasets into distinct groups, proved less effective in identifying consistent and acceptable clusters from a semantic perspective. In contrast, hierarchical clustering demonstrated greater efficacy in finding meaningful groupings. However, the average silhouette score initially provided did not offer clear guidance on the optimal number of clusters, suggesting that the data lacked a strong clustering structure or that the clusters were not well-separated. Consequently, the clustering results were not sufficiently informative, and the groupings did not clearly convey the expected semantic relationships. This indicated that noise and high dimensionality might be obscuring meaningful patterns. To mitigate this issue, Principal Component Analysis (PCA) was employed as a denoising approach. PCA reduced the data's dimensionality while retaining the most significant variance, thereby

[8] Lambda value balances the trade-off between model complexity and goodness of fit. The *lambda.1se* was used for all the regression attempts since such a value tends to produce a more parsimonious model that is less prone to overfitting.



((a)) Screeplot of the percentage of explained variance

((b)) Variance explained by the PCs

Figure 9. Variance captured by PCs

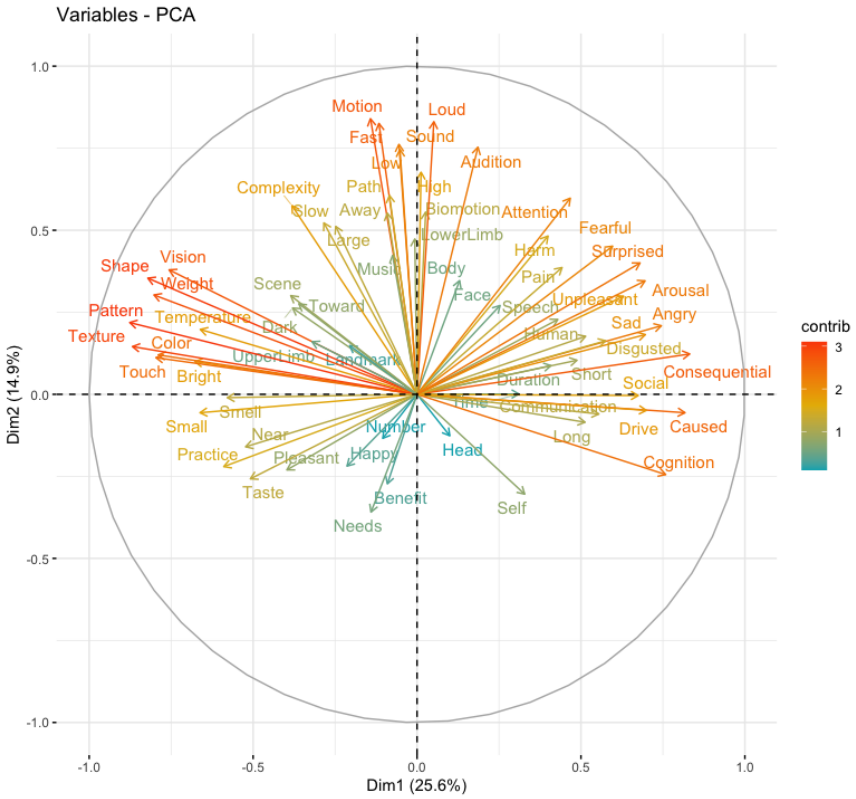


Figure 10. PCs' loadings

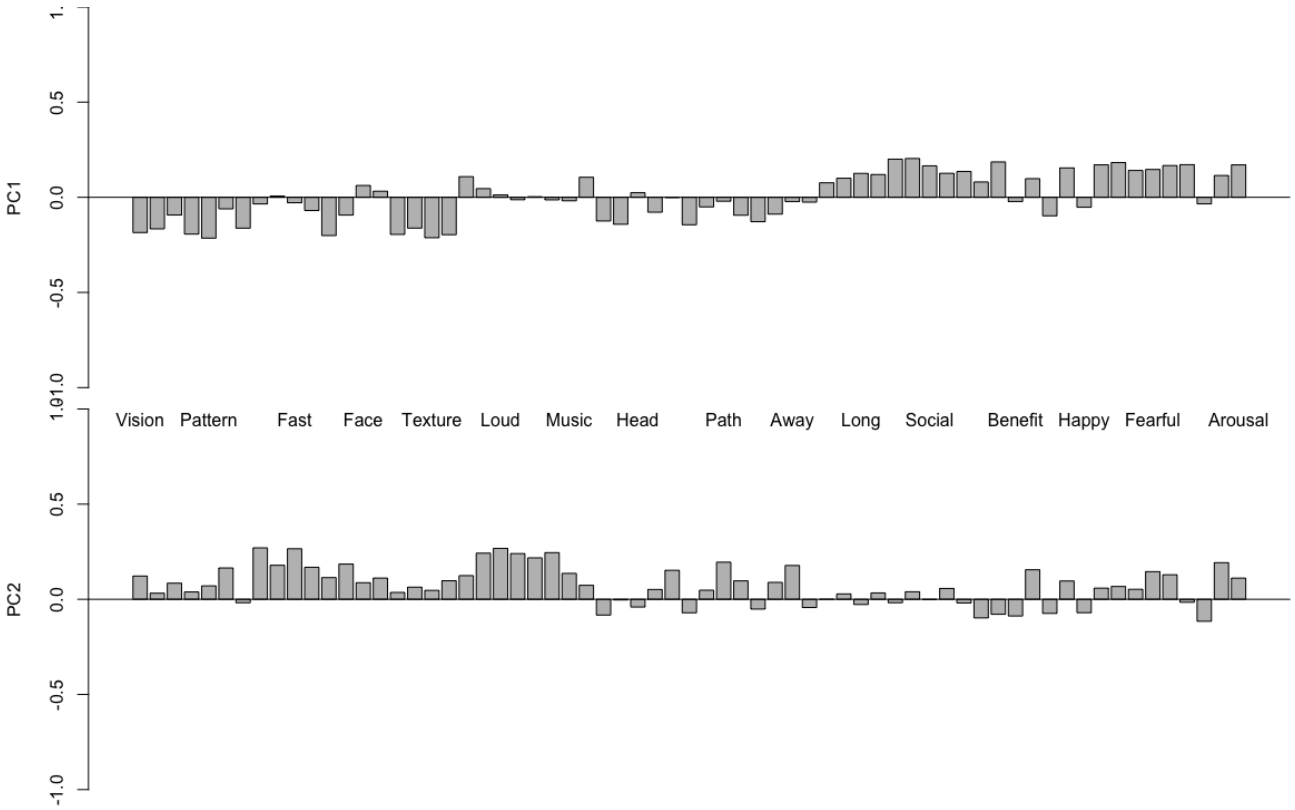


Figure 11. Features contribution on PC1 and PC2

Table 7
Non-zero coefficients used in Lasso regression out of the initial 65 features

Features	*Coefficients*
Vision	0.164964080
Pattern	−0.239966415
Biomotion	−0.187673791
Slow	−0.046411522
Shape	−0.031113976
Complexity	0.502500065
Weight	−0.231003748
Loud	−0.394602592
Music	0.288597551
Speech	0.083764672
UpperLimb	−0.577890160
Practice	0.632002446
Landmark	−0.078626047
Path	−0.086551334
Scene	−0.255575541
Time	−0.007375485
Long	−0.167826216
Short	−0.740172206
Caused	−0.010822897
Consequential	−0.002997683
Human	0.091222033
Self	0.071731126
Cognition	0.158130541
Benefit	−0.809618167

enhancing the clarity of the subsequent clustering analysis. This combination of PCA with clustering techniques provided a more robust framework for uncovering the similarity structures within the data. Despite most of the variance being explained by the first ten components, using the first two proved to be very fruitful. In fact, analyzing the features contributing most to these two dimensions revealed two latent semantic dimensions: *Concreteness* and *Abstractness*. This interesting finding suggested that the dichotomy between concreteness and abstractness could lead to better clustering solutions and raised a further -for now still open- question about their role in defining lexical meaning.

The decision to try biclustering arose from the need to explore every potential method for achieving optimal clustering solutions. Biclustering allowed for the validation of the source data and the semantic information it conveys while also simultaneously performing a feature clustering task. This method proved to be the most effective among the clustering ones, given the high dimensionality of the dataset. The importance of these results lies not only in demonstrating the potential of semantic primitive features to identify lexical items with similar semantic bases but also in retrieving groups of features belonging to the same semantic domain. This underscores the robustness of the biclustering approach.

Similarly, the supervised learning analysis answered the research question positively. The results from the multinomial regression confirmed the validity of this computational approach to lexical semantics and provided valuable insights into the relationship between semantic properties and morphosyntactic classifications.

The successful implementation of the under sampling approach, coupled with feature selection, resulted in a highly accurate classification model that effectively used the semantic information provided by the dataset. This final part of the analysis proved to be particularly interesting because the data that have been used comes from human behaviors. Thus, the positive outcomes in the classification task may indicate that also humans use such semantic information to build more structured knowledge about other language levels, such as the morphosyntactic one. This suggests that latent information within the meaning and their basic components enables humans to distinguish between nouns, verbs, and adjectives. Furthermore, feature selection highlighted intriguing findings. The majority of the features selected and used by the classifier corresponded closely with those that best described the first and second principal components in PCA, reinforcing the hypothesis of a possible important role of the dimensions Abstractness and Concreteness in distinguishing lexical classes and constructing lexical meaning.

7 Conclusion

This study validated the brain-based componential approach introduced and proposed by Binder et al. (2016), focusing on the possibility of identify similarity structures through semantic features and distinguishing morphosyntactic classes using such semantic information. In the unsupervised analysis, despite numerous attempts, some of the clustering methods (i.e., biclustering and hierarchical clustering after applying PCA) demonstrated the effective potential of semantic components in conveying enough information to form consistent groups. This approach's validity was further confirmed by the supervised learning analysis, which revealed a close connection between semantic and morphosyntactic information. This linkage enables classifiers, and humans alike, to distinguish between different lexical classes. In conclusion, it can be said that this study demonstrated the power of such brain-based componential approach in representing and distinguishing semantic categories by the means of the semantics level. Moreover, these findings could represent a good starting point for using such an approach as a support to the development of advanced AI systems capable of understanding and processing language similarly to humans.

Future Directions By confirming the robustness and validity of this brain-based semantic approach, the main scope is to stimulate further empirical and theoretical efforts toward a comprehensive brain-based semantic theory. This means continuing to investigate the possibility of representing meaning through new and more captivated experiential and multimodal semantic features, as well as investigating how these interface with observable reality and human experience. Moreover, the intriguing results regarding the potential of the concreteness-abstractness dichotomy certainly pave the way for further studies to unveil the role that these two interesting latent semantic dimensions play in the construction of meaning and the distinction of lexical items.

Acknowledgements

The author would like to thank Prof. Francesca Chiaromonte at the Sant'Anna University of Pisa for the assistance, the fruitful discussions and support concerning the statistical analysis applied in this paper. This work has been carried out while Davide Testa was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with Fondazione Bruno Kessler (FBK) in Trento.

References

Binder, Jeffrey et al. (June 2016). "Toward a brain-based componential semantic representation". In: *Cognitive Neuropsychology* 33. DOI: 10.1080/02643294.2016.1147426.

Borghi, Anna M et al. (2011). "Manipulating objects and telling words: a study on concrete and abstract words acquisition". In: *Frontiers in psychology* 2, p. 6864.

Bowers, Jeffrey S (2009). "On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience." In: *Psychological review* 116.1, p. 220.

Cree, George S and Ken McRae (2003). "Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns)." In: *Journal of experimental psychology: general* 132.2, p. 163.

Csardi, Gabor (2023). *isa2: The Iterative Signature Algorithm*. URL: https://cran.r-project.org/web/packages/isa2/index.html (visited on 05/2024).

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. DOI: 10.18637/jss.v033.i01.

Gainotti, Guido et al. (2013). "The evaluation of sources of knowledge underlying different conceptual categories". In: *Frontiers in human neuroscience* 7, p. 40.

Harnad, Stevan (1990). "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1-3, pp. 335–346.

Hoffman, Paul and Matthew A Lambon Ralph (2013). "Shapes, scents and sounds: quantifying the full multi-sensory basis of conceptual knowledge". In: *Neuropsychologia* 51.1, pp. 14–25.

Jackendoff, Ray S (1992). *Semantic structures*. Vol. 18. MIT press.

Katja Wiemer-Hastings, Katja and Xu Xu (2005). "Content differences for abstract and concrete concepts". In: *Cognitive science* 29.5, pp. 719–736.

Louwerse, Max M et al. (2005). "The Embodiment of Amodal Symbolic Knowledge Representations." In: *FLAIRS*, pp. 542–547.

Lynott, Dermot and Louise Connell (2013). "Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form". In: *Behavior research methods* 45, pp. 516–526.

McRae, Ken, Virginia R De Sa, and Mark S Seidenberg (1997). "On the nature and scope of featural representations of word meaning." In: *Journal of Experimental Psychology: General* 126.2, p. 99.

Meteyard, Lotte et al. (2012). "Coming of age: A review of embodiment and the neuroscience of semantics". In: *Cortex* 48.7, pp. 788–804.

Rosch, Eleanor et al. (1976). "Basic objects in natural categories". In: *Cognitive psychology* 8.3, pp. 382–439.

Troche, Joshua, Sebastian Crutch, and Jamie Reilly (2014). "Clustering, hierarchical organization, and the topography of abstract and concrete nouns". In: *Frontiers in psychology* 5, p. 81023.