Topological Data Analysis of Multilingual Word Embeddings

Laurel Li[1], Zihe Liu[2], and Jiongli Zhu[3]

{*sil089,zil101,jiz143*}*@ucsd.edu*

1 Introduction

1.1 Main Idea

The inherent diversity and complexity of languages provide a rich landscape for investigation using advanced computational tools. This project aims to uncover the relationships between various languages by analyzing their structural and linguistic features. We employ established techniques from topological data science, an area specializing in the study of space-like mathematical properties, combined with multilingual linguistic analysis. Through these methods, including the use of sophisticated tools such as dendrogram, we strive to enhance our understanding of linguistic interconnectedness and historical development.

1.2 Related Works

Topological methodologies offer a sophisticated framework for analyzing and elucidating the relationships among languages, particularly through the lens of syntactic structure data [PKM19]. These methodologies, rooted in the principles of topological data science, provide valuable insights into the geometric and algebraic connections hidden within language data, thereby allowing researchers to map out complex linguistic relationships in a multidimensional space.

In the realm of natural language processing, the MUSE framework (Multilingual Unsupervised and Supervised Embeddings) plays a crucial role by leveraging multilingual word embeddings to identify and analyze linguistic connections across different languages [CLR$^+$17]. These embeddings represent words or phrases from multilingual sources in a shared high-dimensional space, facilitating the comparison and contrast of linguistic elements across language boundaries.

By integrating multilingual word embeddings with topological analysis techniques, our research addresses a significant gap in the study of lexical and linguistic structures. This innovative approach significantly enhances the granularity of our analysis, enabling the exploration of a broader array of linguistic attributes, including complex lexical relationships often overlooked in traditional studies. Traditional linguistic methodologies typically focus on linear and straightforward analyses and miss the deeper, more nuanced relationships among languages. Our application of advanced topological tools such as persistent homology, mapper algorithms, and multidimensional scaling reveals underlying lexical and structural patterns that are not immediately apparent. This study enriches the understanding of the intricate web of lexical structure, providing new insights into multilingual dynamic complexities and interconnectedness.

2 Methodologies

In this project, we first utilized Mapper to reduce high-dimensional word embedding data from different languages to a lower-dimensional space and constructed a simplicial complex. We then calculated persistence diagrams from the simplicial complex. Using these persistence diagrams, we computed pairwise bottleneck distances and built a distance matrix. Finally, we performed hierarchical clustering on the multilingual data using the Ward linkage method. We first transformed the distance matrix into a condensed form and then applied agglomerative clustering with optimal ordering to generate the clustering results for different languages.

Mapper Mapper is a tool used in topological data analysis to analyze high-dimensional data. It works by mapping the data to a lower-dimensional space using a filter function, which helps to reveal the underlying topological structure. This process defines overlapping intervals in the mapped space, within which a clustering algorithm (such as hierarchical clustering algorithms) is applied to the data points. Each resulting cluster corresponds to a node, and nodes from overlapping clusters are connected by edges. By linking these nodes and edges, Mapper constructs a graph that enables the visualization of the topological structures and patterns within the data, enabling precise analysis of its topological properties.

Persistence Diagram Persistence diagrams [ELZ00] are useful tools for revealing key topological features, such as high-dimensional loops and voids, that appear and disappear as dimensions are scaled within the embedding space. These features are tracked through a process called filtration, where simplices (nodes, edges, etc.) are added incrementally at different levels. In this process, each simplex gets a filtration value indicating when it is included. As the filtration value increases, new topological features appear (birth) and eventually merge or disappear (death). The ”birth” and ”death” times of these features correspond to the filtration values at which they appear and disappear, respectively. These times are represented by points in the persistence diagram, providing an intuitive method to observe the fundamental geometric and topological properties embedded within the languages.

Bottleneck Distance The bottleneck distance [ZC04] measures the maximum distance required to optimally match feature points from one persistence diagram to another. This distance effectively reflects the topological similarities and differences between datasets from different languages. By computing the bottleneck distance, we can quantify the dissimilarity between the topological features of different datasets, providing a rigorous mathematical basis for comparing complex structures. In our analysis, the bottleneck distance was used to create a distance matrix, which served as the input for hierarchical clustering. This approach enabled us to identify clusters of languages with similar topological structures, revealing insights into the multilingual relationships and patterns.

Hierarchical Clustering The hierarchical clustering [Nie16] is a method used to group data points based on their similarities into a tree-like structure called a dendrogram, which reveals its hierarchical organization. In our study, we employed Ward's minimum variance method, an agglomerative approach that minimizes within-cluster variance, creating compact and uniform clusters. It supports our endeavor to uncover and analyze the complex relationships between languages by ensuring the extraction of precise, significant patterns from our linguistic data. This method's robustness and ability to generate detailed dendrograms align perfectly with our goal of enhancing the understanding of multi-linguistic interconnectedness and historical development, facilitating deeper insights into the structural and linguistic complexities across various languages.

3 Experiments

Settings We obtained multilingual word embeddings from an open-source library named FastText [JBM$^+$18, BGJM17]. It provides pre-trained word vectors for 157 languages. However, the size of corpora of different languages are significantly different. For instance, there are only 57 word embeddings of language Afar in FastText, which is much less than 3 million German word embeddings. An incomplete set of word embeddings will lead to inaccurate results, as it only contain a small subset of the vocabulary. To address this, we focused on 44 out of 157 languages, the vocabularies of which are guaranteed to be aligned, ensuring a more reliable analysis.

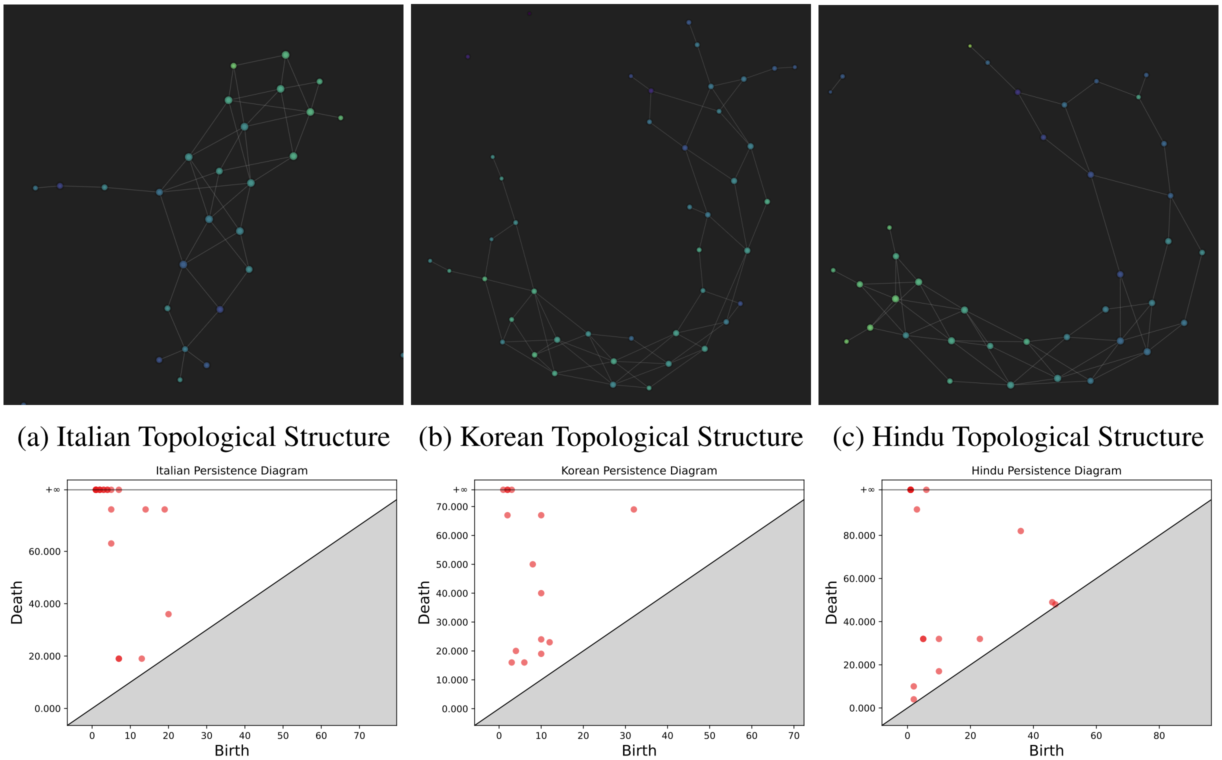
In the experiments, we leveraged the Kepler mapper library [vVSEM19], which implements the mapper algorithm [SMC$^+$07] for topological analysis of high dimensional data. We also utilized the GUDHI library [The15] for generating persistence diagrams. We found that the scalability of the mapper algorithm is poor, resulting in a long runtime when conducting experiments with full vocabularies (did not finish in 6 hours for 1 language). Therefore, we sampled 1000 words from each language to conduct the experiments in this project.

3.1 Linguistic Topological Analysis

Figure 1 shows three examples of topological structures and persistence diagrams we obtained from three different languages including Italian, Korean, and Hindu. By looking at the topological structures, Korean (Figure 1b) and Hindu (Figure 1c) seem to be highly similar, while both of them as dissimilar to Italian ((Figure 1a)). However, visual comparisons might be misleading, as it is highly relevant to the way of visualizing the graphs, and might be influenced by unimportant factors such as the orientation of the graph. In contrast, persistence diagrams serve as a critical way of comparing graphs quantitatively. In fact, the Italian (Figure 1d) and Korean (Figure 1e) are more similar in terms of persistence diagrams, and are less similar to Hindu (Figure 1f).

3.2 Multi-linguistic Analysis

Although persistence diagrams present a powerful way of assessing similarities between languages in a more quantitative or objective way, it is still challenging to compare all 44 languages included in the analysis by looking at their persistence diagrams. To address this, we leveraged the bottleneck distance (can be seen as a pair-wise similarity measure) between persistence diagrams of different languages, and used the bottom-up hierarchical clustering (also known as agglomerative



(a) Italian Topological Structure (b) Korean Topological Structure (c) Hindu Topological Structure

(d) Italian Persistence Diagram (e) Korean Persistence Diagram (f) Hindu Persistence Diagram

Figure 1: Topological structures (a-c) and persistence diagrams (d-f) of different languages.

clustering) to build the dendrogram (Figure 2) that summarizes the similarities between languages. Next, we present three case studies to demonstrate some results we obtained that are consistent with the background knowledge, e.g., findings by linguists.

English vs. French English and French both belong to the Indo-European language family, with English classified under the Germanic branch and French under the Romance branch. Historically, English has been significantly influenced by French, especially following the Norman Conquest of England in 1066, which introduced a large number of Norman French words into English. This historical interaction is reflected in their lexical similarities and shared vocabulary, despite differences in phonology, grammar, and syntax [Ren19]. The dendrogram (Figure 2) reflects a high level of similarity between these two languages, corresponding to the background knowledge of their intertwined histories and lexical borrowing.

Russian vs. Norwegian Russian and Norwegian belong to different language families; Russian is a Slavic language while Norwegian is a Germanic language. Despite these differences, historical and cultural interactions, especially in border regions such as Northern Norway, have facilitated some degree of linguistic exchange and mutual influence. For example, the Russian language has become beneficial in certain Norwegian regions due to economic and cultural ties [Oln17]. In addition, there are bilingual communities where Norwegian and Russian are both used, affecting language acquisition and use [RW17]. Consistent with this background knowledge, the dendrogram

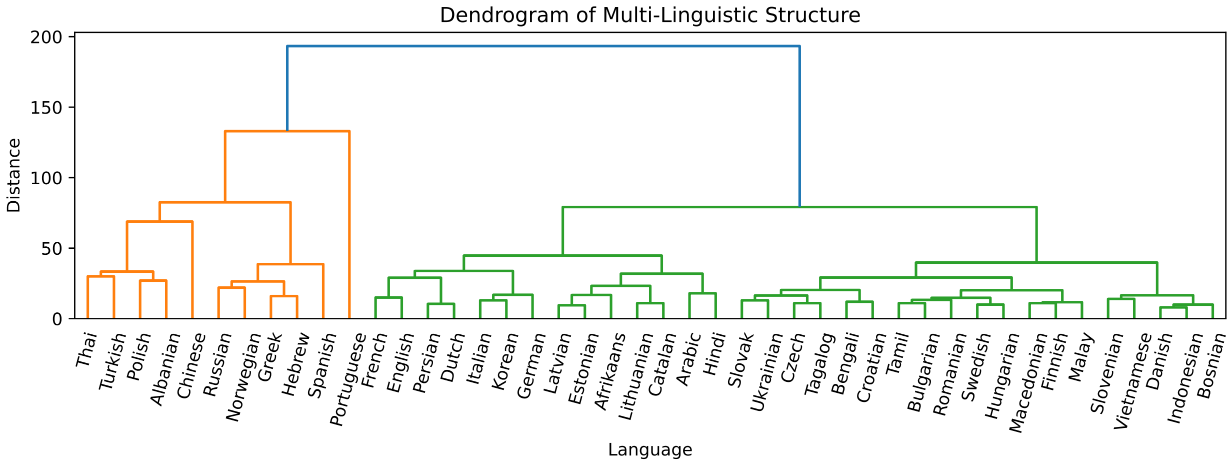


Figure 2: Dendrogram demonstrating the similarities between different languages.

(Figure 2) reflects high similarity between these languages due to these socio-linguistic factors.

English vs. Chinese English and Chinese represent two of the world's major languages, but they belong to entirely different language families; English is part of the Indo-European family, while Chinese belongs to the Sino-Tibetan family [Sha65]. The structural differences are vast, with English being an alphabetic language that uses a Latin script, and Chinese being a logographic language. Additionally, Chinese is a tonal language, whereas English is not. These fundamental differences are also reflected in the dendrogram (Figure 2), where these two languages belong to two distinct and faraway clusters.

4 Conclusions and Future Works

In this study, we have conducted a thorough topological analysis of the structural and lexical properties of 44 different languages using word embeddings. By employing mapper functions, we delineated the complex topological landscapes of each language, which were then quantified into persistence diagrams. These diagrams facilitated a comparative analysis using the bottleneck distance, leading to the development of a hierarchical clustering that uncovered multilingual overarching complexities and interconnectedness.

Our findings corroborate established linguistic theories and demonstrate the utility of topological methods in uncovering multi-linguistic structural and lexical relationships that traditional linguistic analyses may overlook.

In future work, we plan to enhance the scalability of mapper functions to explore larger vocabularies and structural data, and conduct sensitivity analyses on hyperparameters to determine the robustness of linguistic similarities. We will expand our analysis to include a broader range of structural and lexical features, such as syntactic patterns and semantic nuances. Additionally, comparative studies with other linguistic methodologies will be pursued to validate our topological insights. We also aim to apply our methods to the study of language evolution, offering potential contributions to historical linguistics by examining how lexical and structural features of languages have developed and diverged over time.

5 Data Availability

All code and datasets used in this study can be located in our GitHub repository: `https://github.com/RUI2190/Multi-linguistic-Topological-Analysis`.

References

[BGJM17] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[CLR+17] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *CoRR*, abs/1710.04087, 2017.

[ELZ00] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463, 2000.

[JBM+18] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

[Nie16] Frank Nielsen. *Hierarchical Clustering*, pages 195–211. 02 2016.

[Oln17] Margarita Olnova. Russian language in northern norway: Historical, economic and cultural ties. *Russian Journal of Linguistics*, 21(3):587–604, 2017.

[PKM19] Alexander Port, Taelin Karidi, and Matilde Marcolli. Topological analysis of syntactic structures. *CoRR*, abs/1903.05181, 2019.

[Ren19] Vincent Renner. French and english lexical blends¡? br?¿ in contrast. *Languages in Contrast*, 19(1):27–47, 2019.

[RW17] Yulia Rodina and Marit Westergaard. Grammatical gender in bilingual norwegian–russian acquisition: The role of input and transparency. *Bilingualism: Language and cognition*, 20(1):197–214, 2017.

[Sha65] Robert Shafer. The eurasial linguistic superfamily. *Anthropos*, (H. 1./6):445–468, 1965.

[SMC+07] Gurjeet Singh, Facundo Mémoli, Gunnar E Carlsson, et al. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG@ Eurographics*, 2:091–100, 2007.

[The15] The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 2015.

[vVSEM19] Hendrik Jacob van Veen, Nathaniel Saul, David Eargle, and Sam W. Mangham. Kepler mapper: A flexible python implementation of the mapper algorithm. *Journal of Open Source Software*, 4(42):1315, 2019.

[ZC04] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, SCG '04, page 347–356, New York, NY, USA, 2004. Association for Computing Machinery.