THE SHAPE OF WORDS - TOPOLOGICAL STRUCTURE IN NATURAL LANGUAGE DATA

Stephen Fitz
Keio University
Tokyo, Japan
stephenf@keio.jp

ABSTRACT

This paper presents a novel method, based on the ideas from algebraic topology, for the analysis of raw natural language text. The paper introduces the notion of a *word manifold* - a simplicial complex, whose topology encodes grammatical structure expressed by the corpus. Results of experiments with a variety of natural and synthetic languages are presented, showing that the homotopy type of the word manifold is influenced by linguistic structure. The analysis includes a new approach to the Voynich Manuscript - an unsolved puzzle in corpus linguistics. In contrast to existing topological data analysis approaches, we do not rely on the apparatus of persistent homology. Instead, we develop a method of generating topological structure directly from strings of words.

1 INTRODUCTION

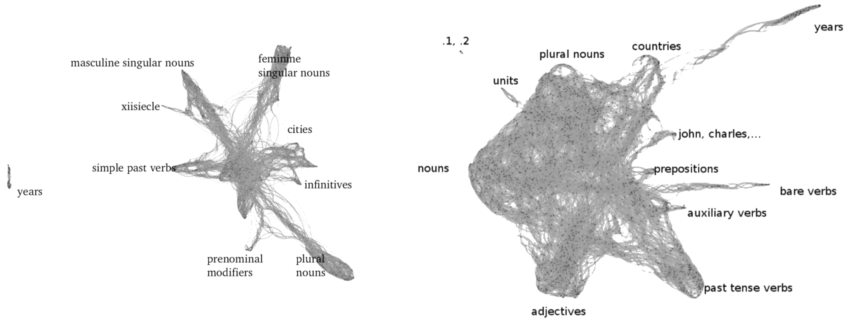

Figure 1: A graph projection of vector space embeddings of word tokens derived from corpora of French (left) and English (right).

The inspiration for the work presented here was an observation made by the authors while working on a separate project involving vector space representations of natural language data. Such techniques are now common in Computational Linguistics and Natural Language Processing. Initially shallow pre-training of early model layers became standard in NLP research through methods such as word2vec Mikolov et al. (2013). Advancements in NLP resulted in progressively deeper hierarchical language representations, such as those derived using self-attention mechanisms in transformer-based architectures Vaswani et al. (2017). Neural language modelling methods, based on techniques such as BERT Devlin et al. (2018), use vector space representations of linguistic units to predict words based on context found in raw text corpora. Interestingly, word vectors obtained through such embedding methods do not fill the ambient space uniformly, but rather form lower dimensional structures within it. Figure 1 shows a graph projection of word embeddings into an ambient vector space obtained from articles in French and English. The embedding method uses word contexts to bring vector representations of words that appear in similar contexts close together. While looking at those projections, it is hard to ignore the extreme differences in shape between languages. The work presented here aims at studying a notion of shape in the context of natural language data, by developing theory and algorithms for probing linguistic structure from a topological viewpoint. Subsequent research by the authors aims to investigate links between topological aspects of natural language data and its representations in vector spaces, which are at the core of recent developments in the field of NLP.

The predominant method of deriving linguistic unit representations within modern language models takes advantage of self-supervised techniques of masked sequence training. This amounts to gradient descent optimization, resulting in representations that are informative towards predicting the contexts in which words appear within the corpus. Because of this, there must be a natural relationship between the topology of raw text data (as expressed by word co-occurrence patterns), and the shape of the resulting embedding manifold (as expressed by the distributions of word vectors). In order to study natural language on the side of LM based vector space embeddings from a topological perspective, we can use existing tools of persistent homology. These techniques take advantage of the fact that such linguistic unit representations exist in a metric space, which naturally comes equipped with the open ϵ-ball topology, and its inner product encodes linguistic structure due to the autoregressive model training process. For this reason, topology based on proximity in the embedding space must hold a natural relationship to the word and context topologies based on word co-occurrence patterns. Among the techniques we can use here, are topological persistence modules from a Vietoris-Rips complex filtration. We would like to understand the relationship between those vector space representations and the raw text data used to induce them. Although we have potential tools to study the vector representations from this perspective, there is currently no analogue on the side of raw text data. This paper develops a method of analyzing raw text data from a topological perspective. Using such representation, we can then study the relationship between neural language representations and raw text data in a uniform fashion, by mapping both into the category of topological spaces and continuous maps. The reason we can not apply the same techniques of persistent homology on the side of the corpus directly, is because natural language data is in form of discrete tokens of words, which do not come with a canonical notion of topology such as that on the side of the embedding. However, literature in pure mathematics provides theoretical foundation form which we can build algorithmic solutions to this problem.

Over the past century, we saw an emergence of new branches of Mathematics exploring the structure of high dimensional objects. These ideas came initially from the Erlangen Program outlined by Felix Klein in his seminal work on the formalization of geometry as the study of invariants under algebraically defined groups of transformations Klein (1872). The body of research produced through this program led to the development of category theory as a unifying language connecting previously isolated branches of Mathematics. A particularly powerful type of such category theoretical relationship is expressed by the homotopy and homology functors linking the realms of topology and abstract algebra. These notions belong to the branch of Mathematics known as algebraic topology, which has been in accelerated development over the past century, producing a set of powerful tools for probing global and local properties of manifolds.

The first goal of our project is to build a bridge between topology and language. Such a connection will then allow us to adapt powerful machinery developed over the past century in pure Mathematics for the analysis of topological manifolds, and employ it for the study of linguistic structure encoded in corpus data. In particular, we are concerned with the development of algorithms for associating topological manifolds directly to raw corpus data. The requirement for these algorithms is that the topology of the resulting manifold encodes word co-occurrence patterns, and that no extrinsic information is imported in this association process. The distributional hypothesis, which forms a fundamental assumption behind neural language model training (best expressed in words of John Rupert Firth - ”you shall know a word by the company it keeps”), states that syntactic and semantic relationships between words can be inferred from their context (i.e. co-occurrence patterns with other words in the corpus). We use this idea as a basis for the construction of a simplicial complex, where 0-dimensional cells are identified with word tokens, and higher-dimension features are determined by the n-gram patterns within a given corpus of text. The resulting topological structure will be called the *word manifold*, and exploration of its properties is one of the primary goals of this work. The *word manifold* is a topological space obtained directly from raw natural language data, without the use of any neural encoders. It is a topological manifold, and it does not need a metric space embedding. Instead, the topology expressed by its simplicial complex structure, encodes word-context relationships directly.

Gunnar Carlson et.al. applied topological data analysis to patches of pixels from naturally occurring images Carlsson et al. (2008). The analysis led to a conclusion that the shape of the image manifold under study could be approximated by a Klein bottle. This realization led to a novel compression algorithm for images taking advantage of a parametrization of the pixel space that mapped 3x3 patches of images onto points on a sub-manifold homeomorphic to the Klein bottle within the image manifold. To the best of our knowledge such approaches have not been applied in the field of Computational Linguistics. Understanding the topological structure of natural language representations can lead to reparametrization techniques, allowing for significant model compression. Our current research efforts aim to develop methods for deriving more parameter-efficient language modeling techniques, by studying the topological structure of natural language data, which can then be used to reparametrize the embedding or introduce topological regularization techniques into the field of NLP.

2 The Word Manifold

2.1 Skeleta

We start with a corpus of natural language text $\mathcal{C}$, and a window size k, which defines the maximum n-gram size under consideration. The first step in constructing the *word manifold*, is the induction of what we will call *skeleta*. The n-skeleton S_n is the collection of all (n+1)-gram simplices $[w_0, w_1, \dots, w_n]$ for which the following condition is satisfied. Every subsequence, of any length, composed of words from $\{w_0, w_1, \dots, w_n\}$ appears somewhere in $\mathcal{C}$ within a window of k word tokens. Observe that the 0-skeleton S_0 is generated by the lexicon extracted directly from $\mathcal{C}$. Furthermore, the $(k-1)$-skeleton (top dimensional cells of the complex) will be composed of k-grams found in the corpus. Note that the collection of all skeleta defined this way on the n-grams of the corpus, forms an abstract simplicial complex. Furthermore, the simplices encode the distribution of words within the predefined n-gram window size, and the orientation of simplices expresses the order of words as they appear in the sentences of the corpus.

2.2 Boundaries

Once all skeleta are induced from the corpus, we define the *boundary matrices* in the following way. First, we impose alphabetical ordering on the 0-skeleton. Then the ordering for higher dimensional skeleta follows from the canonical ordering on the cartesian product (i.e. an n-gram is ordered as if it was an element of the cartesian product of n copies of the lexicon). We then define a n-indexed collection of matrices $\mathcal{B} := \{B_n | n \in \{0, .., k-1\}\}$, where k is the maximum window size, as in the previous subsection. We will call the elements of $\mathcal{B}$ the *boundary matrices*. They are defined in the following way. The n-dimensional boundary matrix $B_n \in \mathcal{B}$ is a $r \times c$ matrix, whose columns correspond to elements of the n-skeleton induced from the corpus in the previous step, in the ordering defined above. The columns of B_n are identified with the n-skeleton, while rows correspond to the elements of the $(n-1)$-skeleton in the ordering defined above. That is $c = |S_n|$ and $r = |S_{n-1}|$. The entries of the *boundary matrix* B_n are then defined as follows. For each column, we look at the corresponding n-simplex $s \in S_n$, and define its formal boundary ∂s by the following formula, where hat designates that the term under it was removed.

$$\partial[w_0, \dots, w_n] = \sum_i (-1)^i [w_0, \dots, \hat{w_i}, \dots, w_n]$$

We extend this definition linearly to the entire free group generated by the skeleta. This defines a group homomorphism between free abelian groups generated by S_n and S_{n-1}. The *boundary matrix* is then a matrix representation of this group homomorphism - that is the entries of B_n are the coefficients $(-1)^i$ in the above formula. The resulting collection of matrices $\mathcal{B}$ will then be used in computations to analyze the topological structure of the *word manifold*. The *boundary matrices* obtained from natural language corpora are highly sparse tensors with entries in $\{-1, 0, 1\}$, which allows for efficient computations using BLAS algorithms.

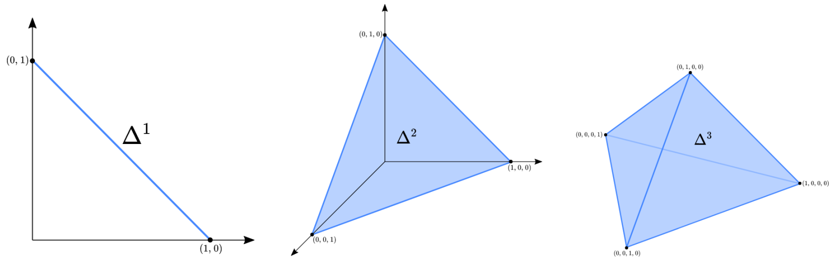

Figure 2: Visualization of a simplex in different dimensions. The algorithm we develop in this paper maps a corpus of natural language text onto a simplicial complex. Simplices can be represented geometrically as convex hulls of standard basis vectors in $\mathbb{R}^n$. In the case of the *word manifold* the 0-simplices (i.e. vertices) correspond to the lexicon, and higher dimensional simplices come from n-grams of words found in the corpus. The gluing together of simplices into a topological space, as well as the definition of the boundary operator is determined by co-occurrence patterns of words in context.

2.3 Homology

Homology theory constructs functorial mappings from the category of topological spaces and continuous maps to the category of abelian groups and group homomorphisms. In this translation homotopy equivalent manifolds induce isomorphic homology groups. This allows us to employ powerful theorems from commutative algebra for the study of topological spaces. Details of this construction go beyond the scope of this paper, but the reader can find a good introduction in Hatcher (2001). The resulting method allows us to generate a sequence of free abelian groups together with group homomorphisms, known as a chain complex. In our case, the groups are generated by the n-grams of the corpus, and we will call them *context groups*, denoted by C_n. There is a natural homomorphism between the context groups in subsequent dimensions, which maps a given n-gram to a formal sum of (n-1)-grams, as explained in the previous subsection. The collection of all the context groups with these n-gram homomorphisms forms a chain complex.

$$\cdots \to C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \to C_1 \xrightarrow{\partial_1} C_0 \to 0$$

The groups C_i above are associated to the corpus by the skeleta induction procedure - that is, each skeleton is a generator set for the corresponding context group. Homology theory quantifies the structure of *cavities* (*holes* in different dimensions) in the topological space under consideration. There are many possible procedures resulting in equivalent homology theories (i.e. the resulting homology groups are isomorphic). For the purposes of this paper, we employ a combinatorial version of this theory known as simplicial homology (see figure 2). The key property of the context group homomorphism $\partial_\bullet$ above is that $\forall_n \partial_{n-1} \circ \partial_n = 0$, which implies that $\mathrm{im}\partial_n \subseteq \ker \partial_{n-1}$. The elements of $\ker \partial_\bullet$ are called cycles, and the elements of $\mathrm{im}\partial_\bullet$ are called boundaries. Homology of the *word manifold* is then defined as the group theoretic quotient of cycles modulo boundaries.

$$H_n(C_\bullet, \partial_\bullet) = \frac{\ker \partial_n}{\mathrm{im}\partial_{n+1}}$$

These groups form algebraic invariants of topology, so homeomorphic word manifold structures will result in same groups up to isomorphism. We are often interested not in the particular groups but just in their ranks, known as Betti numbers, which can be interpreted as counting cavities in all dimensions of the manifold, and thus give a coarse description of its shape (see figure 3 for an example). The following proposition allows us to efficiently compute the Betti numbers using basic linear algebra subprograms.

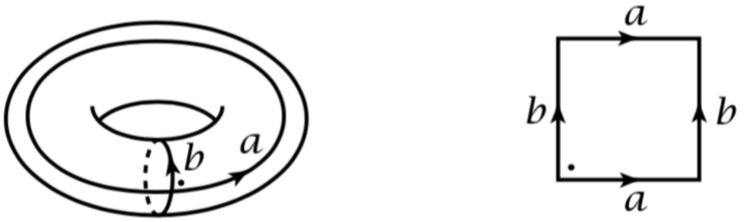

Figure 3: The torus has a single connected component which corresponds to a single 0-dimensional ”hole”. It has two 1-dimensional holes - one measured by the class of loops wrapping around the main circle (represented by a) and another by the class of loops going through the middle hole (represented by b). These loops are independent because there is no way to continuously deform any of the loops in the first class into any of the loops in the second class. Finally, the torus has a single 2-dimensional hole, which is generated by the tire shaped cavity inside its surface. These numbers of independent holes are the *betti numbers* of the torus. Right side shows a cellular decomposition, which yields the torus by taking a quotient according to the labels given.

Proposition: Let A be an $m \times n$ matrix, and B be an $l \times m$ integer matrix with $BA = 0$. Then

$$\ker(B)/\mathrm{im}(A) = \bigoplus_{i=1}^{r} \mathbb{Z}/\alpha_i \oplus \mathbb{Z}^{m-r-s}$$

where $r = \mathrm{rank}(A)$, $s = \mathrm{rank}(B)$, and $\alpha_1, \ldots, \alpha_r$ are the nonzero elements of the Smith normal form of A.

After interpreting our chain complex as a sequence of finitely generated abelian groups, where the generators correspond to the n-grams of words extracted from the corpus, and computing the Smith normal form of the boundary map, we can extract the number of cavities in each dimension of the *word manifold* simply by reading the nonzero columns of the diagonal matrix in the resulting decomposition.

In the results section we present comparison of 6 languages on corpora of 200k sentences per language extracted from news articles on the same topics, during the same period of time. Because the chosen corpora are semantically similar, the differences between word manifold topologies can be attributed to inherent linguistic information encoded in the data. Additionally, we perform studies with algorithmically generated data, and an analysis of an unknown script.

3 Results

We performed a total of 1320 experiments with eight corpora of natural language text. Each experiment involves a random sampling of 5000 sentences. The sampling for each experiment was repeated 10 times, and the results were summarized with a box plot. The experiments took about a month on a single Ryzen TR node with 48 CPU cores, 256 GB of RAM, and an NVIDIA RTX 3090 GPU. The first experiment derives word manifolds from a smaller test corpus constructed manually by linguists at Brown University Francis & Kucera (1979). In order to ablate the contribution of linguistic structure in the formation of topological features within the word manifold, we also construct synthetic corpora, obtained from this corpus by procedural generation targeting specific effects. This artificially generated data is compared to natural data in the following three experiments. The next six sets of experiments target larger corpora of modern languages obtained by crawling native news sources over the same period of time: Arabic, English, French, German, Japanese, Russian. Finally, in the eighth set of experiments, we apply the same analysis to transliterations of the Voynich Manuscript, shedding new light on this unsolved puzzle in corpus linguistics.

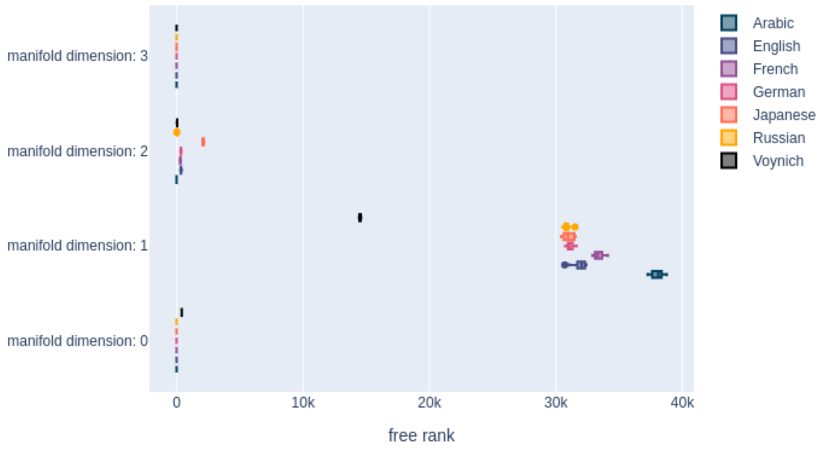

Figure 4: Comparison of topological structure in different dimensions of the *word manifolds* induced from 6 modern languages, and a manuscript of unknown origin.

3.1 Modern Languages and the Voynich Manuscript

Figure 4 shows box plots comparing free ranks of homology groups in the first four dimensions of the *word manifold*. Those results can be interpreted as a coarse summary of the structure of cavities in different dimensions. First observation is that dimensions 1 and 2 seem to be most sensitive to grammatical differences between languages. Voynich is a clear outlier in dimensions 0 and 1. This is interesting, as the experiments with synthetic data in the next subsection suggest that holes in dimension 0 are related to parts of speech, and holes in dimension 1 encode sentence level syntactic information. In dimension 1, French has more topological structure than English, while in dimension 2 English exhibits more complex features than French. This is interesting when looking at visualization of their vector space embeddings in figure 1, where English appears more ball like (which would make sense, since dimension 2 is defined by sphere like objects)! The study of topological relationship between *word manifolds* associated to raw natural language text, and their vector space embeddings induced by language models is the topic of our current work in progress on this subject. The manifold with the closest match in dimension 1 with the Voynich Manuscript is Russian, which would align with the known history of the manuscript, providing support to several mainstream theories about its possible origin. Japanese is a clear outlier in dimension 3, manifesting nontrivial topological structure. This is interesting since it is the only Asian language included in the study, suggesting that higher dimensions of the *word manifold* possibly measure differences between language groups. It is a topic we plan to investigate in future work.

3.2 Synthetic Data

In this study we took a corpus of natural language text in English, and used it as a basis for generating additional corpora targeting specific structures. These corpora were constructed from the natural language text by means of three transforms, aimed to be progressively more destructive to linguistic structure: permutation, zipf, and uniform. The permutation transform simply permutes the order of words within each sentence of the corpus. This operation preserves many word interactions, limiting possible context of any given word. The zipf transform generates sentences randomly by unigram sampling form the Zipf's Law distribution. Finally, the uniform transform samples words uniformly at random. Both zipf and uniform transforms maintain the natural distribution in sentence length. We then derived the *word manifolds* from both the natural and synthetic corpora and compared them against each other. Figure 5 shows box plots of free ranks of homology groups of

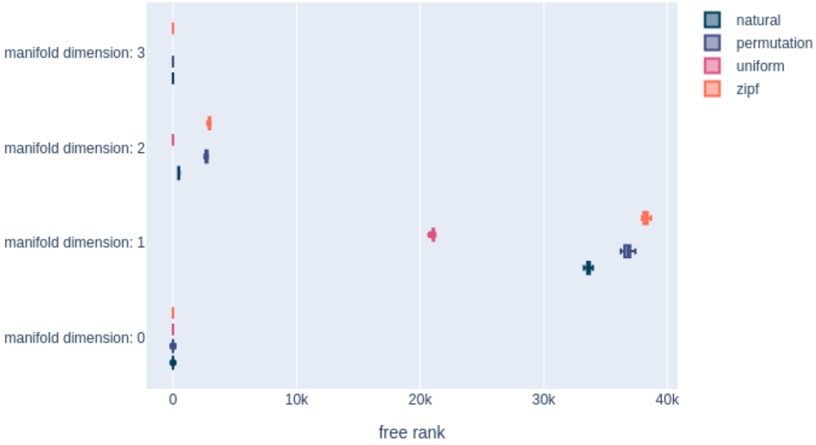

Figure 5: Comparison of natural and synthetic corpora.

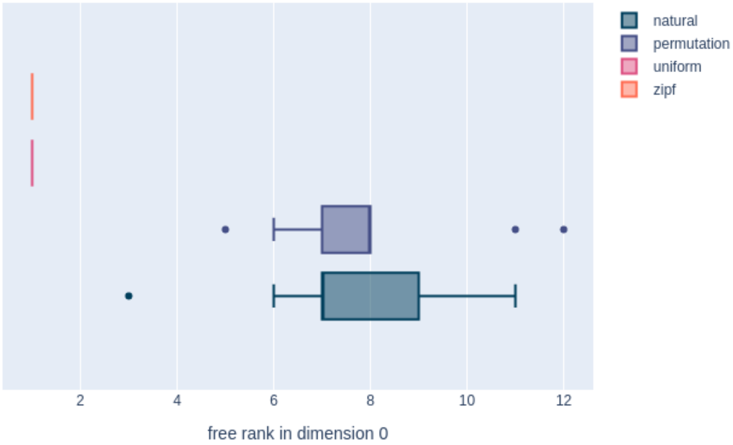

Figure 6: Dimension 0 is possibly related to the parts of speech!

those *word manifolds* associated to natural, and three synthetically generated corpora. We observe significant differences in dimension 1 and 2. In dimension 1, the uniform corpus is an outlier, with significantly less topological features. This suggests that dimension 1 is at least partially associated with sentence structure, as uniform transform destroys sentences. On the other hand, permutation and zipf transforms produce too many holes in dimension 1, which suggests that dimension 1 must also encode syntax and semantics to some degree, going beyond simple word statistics. Interestingly, dimension 0, which measures the number of connected components in the manifold, is possibly related to the parts of speech. Figure 6 shows a zoomed in view of the zeroth dimension. We note that the natural corpus seems to have around 8 connected components, which would correlate with 8 classical parts of speech Thrax (2nd century BC). The permutation corpus also exhibits nontrivial

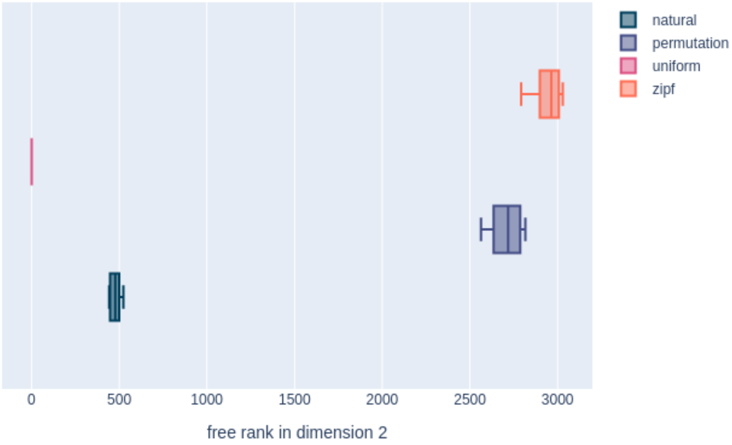

Figure 7: Dimension 2 clearly separates natural from synthetic data.

homology in this dimension. By contrast the uniform and zipf transforms produce fully connected *word manifolds*. This makes sense, since the permutation transform still retains sentence structure to some degree, limiting possible word interactions, while the other two transforms completely destroy sentence boundaries, which allows for arbitrary connections between words to be formed. Figure 7 shows summary of dimension 2. In this dimension the permutation and natural corpora are the closest, and clear outliers from the rest. This suggests that higher dimensions of the *word manifold* go beyond statistical correlations between words and involve higher level linguistic structure. Recall that the uniform and zipf corpora are generated using simple unigram distributions, which destroys most of grammatical structure. By contrast the permutation transform will retain a lot of n-gram context data, which contains syntactic and semantic information about words. This is clearly manifested in the 2-dimensional cavity structure of the associated *word manifolds*.

4 Conclusion

We presented a novel approach to the study of linguistic structure through a topological lens. We defined the notion of a *word manifold*, and described an algorithm for its induction from raw natural language text. We also discussed the computation of algebraic invariants of the word manifold in form of homology groups. We then applied our method to a variety of languages, synthetically generated data, and an unknown script from the 15th century, currently held at Yale University library. These results show that the topology of the *word manifold* is influenced by linguistic structure expressed by the corpus. Furthermore, we can interpret dimensions of the *word manifold* by comparing natural and synthetic data.

In future work, we plan to explore the relationship of the word manifolds associated to raw text data, and vector space representations of linguistic units that arise in neural language models.

References

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *International journal of computer vision*, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

W Nelson Francis and Henry Kucera. Brown corpus manual. *Letters to the Editor*, 1979.

Allen Hatcher. *Algebraic topology*. Cambridge University Press, 2001.

Felix Klein. *Vergleichende betrachtungen über neuere geometrische forsuchungen*. A. Deichert, 1872.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. 2013.

Dionysios Thrax. Art of grammar. 2nd century BC.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.