




## Research Article

# A Language and Its Holes: The First-Order Homology of the Large-Scale Geometrical Structure of a Natural Language

Vasilii A. Gromov , Quynh Nhu Dang , and Asel S. Erbolova 

Laboratory of Analysis of Semantics, Centre for Language and Semantic Technologies, Faculty of Computer Science, HSE University, Moscow, Russia

Correspondence should be addressed to Vasilii A. Gromov; [stroller@rambler.ru](mailto:stroller@rambler.ru)

Received 4 September 2024; Revised 4 September 2025; Accepted 8 October 2025

Academic Editor: Pramita Mishra

Copyright © 2025 Vasilii A. Gromov et al. Complexity published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

The present paper employs topological data analysis methods to reveal ‘holes’ (stable persistent homologies) in the semantic spaces of words, bigrams, and trigrams of the English and Russian languages, and to ascertain their boundaries. Furthermore, the paper selects those holes that belong to the large-scale (coarse-grained) structure of the language that are not just local inhomogeneities of the sample—it appears that there are around a dozen of them for each of the languages (English and Russian). These boundaries delineate ‘blind spots’ of the respective language—the regions of the semantic spaces that do not contain words/bigrams/trigrams of the language—that is, regions of concepts that the language cannot see through its lens. The secondary goal of the paper is to solve the bot-detection problem in its strong statement, that is, one trains the classifiers on one set of bots and tests on the another set of bots. To this end, we estimate the average distances from words, bigrams, and trigrams of a text to the boundaries of the nearest ‘hole’, for texts both written by humans and generated by bots, and construct classifiers. The classifiers show comparatively good results: the average accuracy amounts to 0.8.

**Keywords:** a natural language; bot detection; large-scale structure; persistent homology; topological data analysis; ‘holes’ in a language

## 1. Introduction

Gromov and Migrina [1] argue that a natural language is a (self-organised-critical) unified system. Modern methods to embed words and  $n$ -grams of a natural language into the ( $d$ -dimensional) Euclidean space (to construct embeddings) [2] raise the question about geometrical properties of this unified system [3]: Gromov et al. propose to examine, for a given natural language, a set of embeddings for all words ( $n$ -grams) of a language. Assumedly, the set of all observed words ( $n$ -grams) constitutes the representative sample of points that belong to the  $d$ -dimensional surface (more precisely, the fractal geometrical object—the language fractal

structure)—thereby one can examine this sample in order to ascertain properties of this surface, the language fractal structure. The authors propose to use the term *Hailonakea* (‘the sign immensity’ from Hawaiian) for the set of all language fractal structures for  $n = 1, 2, \dots$  (A reader may click the link [https://en.wikipedia.org/wiki/Laniakea\\_Supercluster](https://en.wikipedia.org/wiki/Laniakea_Supercluster) in Wikipedia to see a photograph of the Laniakea Supercluster, the largest object of the observable Universe. Each pixel of the photograph corresponds not to a galaxy but to a cluster of galaxies; the photograph gives an insight into the coarse-grained structure of our Universe).

The present paper examines the topological properties of the language fractal structures for  $n = 1$  (words),  $n = 2$

(bigrams), and  $n = 3$  (trigrams) for the Russian and English languages. To run the simulations in order to estimate the intrinsic dimensions, we employ the corpora of the literature for the Russian and English languages. The reasons why we chose these very languages are as follows: (1) both languages enjoy a large corpora of literature texts; (2) they are inherently different, in a sense: The Russian language possesses rich inflections and free word order [4], whereas English is an isolating language, and it tends to follow strict rules on word order [5]. The authors believe, maybe somewhat quaintly, that the literature constitutes a core of the respective language. Consequently, we think that it is quite reasonable to use  $n$ -grams extracted from the corpora of literary masterpieces in order to investigate the respective languages.

The secondary goal of this paper is to solve the bot detection problem, that is, to distinguish texts written by humans and those generated by bots. More to the point, as in [6], we use the problem statement that distinguishes texts of all bots and all humans (not the texts of a particular bot(s), as in the conventional problem statement). Namely, we test the null hypothesis that texts written by humans contain more  $n$ -grams close to the boundaries of the ‘holes’ than those generated by bots: humans tend to produce more unexpected sequences of words than bots.

The rest of the paper is organised as follows. Section 1.1 reviews the related works, the papers that serve as a starting point for the present one. Section 1.2 introduces the problem statement. Section 1.3 discusses the methods employed; Section 1.4 reveals the homologies for Russian and English fractal language structures. Section 1.5 presents the results for the bot detection problem. Finally, Section 2 presents conclusions and future directions. The code and resulting data are available at the repository: <https://github.com/qynhu-d/stb-tda>.

*1.1. Related Papers.* It seems to be helpful to approach the available literature from two angles. First, we review papers that examine an entire language as a natural-science object standpoint (apparently, this goal drives N. Chomsky to establish several variants of generative grammar [7]); then, those that apply topological data analysis (TDA) to reveal ‘holes’ in a given set of points. Importantly, this paper relies on the belief that a natural language constitutes an array of meanings (concepts), rather than that of words (symbols), and one should treat it as a semantic space (space of meanings). Unfortunately, most papers that study a natural language as a whole concern themselves with ‘words’ rather than ‘meanings’. Nevertheless, we may indicate several monographs [8–10] that explore ‘words’ in order to reveal language process properties for ‘meanings’. Gromov and Migrina [1] examine a natural language as a whole unity; the authors prove that a natural language constitutes a self-organised critical system, with a literary masterpiece (or any other text) constituting a power-law governed ‘avalanche’ in a semantic space; Gromov et al. [3] estimate intrinsic dimensions of the language fractal structures. Tanaka-Ishii [9] analyse long correlations for the English and Japanese

languages. Due to the nature of the Japanese language, one can establish nearly one-to-one correspondence between hieroglyphs (kanjis) and meanings. Consequently, one can cautiously extend conclusions made for hieroglyphs to elements of a semantic space (‘meanings’). The language demonstrates long correlations, those of words placed 10–15 positions apart (this significantly exceeds the length of a typical  $n$ -gram in natural language-processing procedures). Dębowski [8] investigates the power laws of a natural language: He demonstrates that this very class of distributions governs most fundamental language processes.

As we treat a text, written or spoken in a natural language, as a unified mathematical object (to be studied), the ‘avalanche’ in terms of self-organised criticality theory, we can treat a sequence of word embeddings of a literary masterpiece a unified mathematical object too:  $d$ -dimensional time series. Gromov and Dang [11] propose to use the term ‘a semantic trajectory of a literary masterpiece’ for such a time series; they also prove that most semantic trajectories are chaotic time series—this allows studying these mathematical objects in the frameworks of the theory of complex systems. The paper also reveals that languages of various language groups differ in chaoticity (incidentally, this constitutes a criterion to assess the quality of translation, both human and machine). The papers [12, 13] explore semantic trajectories in order to ascertain properties of the strange attractors of the dynamical systems that generate these trajectories. In general, a set of all semantic trajectories (for instance, for all pieces of the national literature corpus), that is, either various trajectories of a single dynamical system (under this assumption, one should talk about the strange attractor of a given natural language, with its minimum inertial manifold containing the set of all word [ $n$ -gram] embeddings) or trajectories of various dynamical systems (under this assumption, one should talk about the strange attractor of a given literature masterpiece), which move along the points of the respective language fractal structure, betrays the geometrical characteristics of the structure itself (one can apply these characteristics, both dynamical and geometrical, to solve the bot detection problem [6]. Interestingly, the results for the spaces of  $n$ -grams appear to be much more fruitful than those for the space of words—bots, if anything, use the same dictionaries as humans do; however, humans tend to produce much more unexpected sequences of words [meanings] than bots do).

Another possible line of attack on the language as a whole is complex network models. For instance, Garg et al. [14, 15] investigate the dynamics of natural language processes by means of graph models of a language. Stanisiz et al. [16] also discuss linguistic networks and how they reveal nonrandom, hierarchical structures useful for semantic analysis, highlighting language-specific organisational principles. The authors also consider how written texts exhibit long-range correlations, fractal and even multifractal patterns, particularly when analysed as time series based on sentence or punctuation-segmented phrases. In the present paper, we examine the embedding space of Russian and English languages. We examine formal fractal dimensions in

another study [3], in which intrinsic dimensions of language fractal structures are estimated.

In the rest of the section, we discuss the methods and applications of TDA. One of the main advantages of TDA lies in its applicability to complex systems [17–21]. In cosmology, Bermejo et al. [22] reveal structural patterns in the cosmic web. Skaf and Laubenbacher [23] review the applications of persistent homologies in biology and medicine. They discuss the application of TDA methods in ‘clinical care and precision medicine, medical images analysis, medical diagnostic accuracy, biological research (‘omics’ sciences), structural biology, immunology, epidemiology and many more’. Meng et al. [24] examine various DNA structures by means of localised weighted persistent homology. It appears that topological properties are frequently efficient to cluster various data. Dey and Mandal [25] use filtrations of simplicial complexes to model hierarchical structure of protein molecules; Corcoran and Jones [26] to reveal voids in geographical data and Caputi et al. [27] and Yoo et al. [28] to examine brain connectivity.

In order to study such a complex system as a natural language, one can also apply TDA methods.

First pioneering studies in this field started back in 1960–1970s [29, 30]. The topology of language spaces using density-based approach was studied in [31]. In this work, the authors focused on relations within sets of languages. After establishing the equivalence relations between languages, the quotient spaces were examined in terms of general topology properties. All these studies were large scale, focusing on a language as a single point, sometimes not considering algebraic topological properties of an individual language.

In [32], the authors emphasise the importance of systematically categorising and organising traditional Chinese cultural elements to enhance their extraction and contemporary relevance, utilising the Taiping Imperial Encyclopedia as a primary data source. Employing unsupervised word segmentation and the TF-IDF algorithm, the research establishes a two-dimensional orthogonal classification system for cultural topics, revealing a scale-free network structure with significant community and hierarchical features, where the top 12 identified communities constitute 91.77% of the network. In [33], a latent geometric framework for understanding how humans retrieve knowledge through semantic networks is proposed, addressing the challenges of observing the underlying topology of concepts. This framework not only distinguishes between healthy individuals and dementia patients but also aims to enhance assessments of neurodegenerative diseases and inform targeted nonpharmacological therapies.

Zhu [34] uses filtrations to examine sequences of words to classify verses (in order to compare those written by children and grown-ups). The author introduces the similarity filtration with time skeleton (SIFTS) algorithm, which applies persistent homology to natural language processing by identifying semantic ‘tie-backs’ in text documents, offering a novel representation of document structure, illustrated through various texts including nursery rhymes and adolescent writings. Elyasi and Hosseini Moghadam [35] use persistent homology and mapper methods to classify

masterpieces of Persian poetry. Savle et al. [36] derive topological features in order to classify documents for a legal entailment task. They combined persistent homology features with TF-IDF and found that such augmentation is informative—F-score is improved by 14%. Tymochko et al. [37] reveal ‘holes’ in scientific papers to identify discrepancies between titles and abstracts. The authors employ TDA techniques to identify logical and literary gaps in research papers, demonstrating that the integration of topological features with natural language processing and time-series analysis significantly enhances the detection of fraudulent papers. In [38], the conceptual space analysis of a language is carried out. Authors focus on the space as a complex network and study its small-worldness, shortest paths and other network-related properties.

The approach to finding topological properties derived from language was also proposed in a more recent study [39], in which the authors create a so-called ‘word manifold’ encoding the grammatical structure of the corpus of texts. The authors propose a discrete version of Vietoris-Rips filtration obtained from the weight of  $n$ -grams. Authors note that their topological approach at different dimensions distinguishes the synthetic data from real language (dimension 2), or different parts of speech (dimension 0), or nontrivial topological structure in Japanese at dimension 3.

Many papers [35, 40] concern themselves with the interpretability of TDA, its obvious strong point. Nevertheless, to the best of the authors’ knowledge, nobody has applied TDA methods to a natural language as a whole—the research presented above focus on texts from one domain only. In the present paper, the authors examine the entire language (the English and Russian ones) by observing set of words, bigrams and trigrams to reveal ‘holes’ (stable persistent homology classes) in its semantic space. We construct a large-scale embedding space, showing the precise placement of each word/bigram/trigram in each language and create a complex and high dimensional representation. We then derive features based on the distances to the nearest holes for each language, contrary to previous research where topological features are extracted for each text separately.

**1.2. Problem Statement.** In order to ascertain the topological, coarse-grained structure of a natural language, one should

- For a given set of texts of a given natural language  $\mathfrak{S} = (\Omega_1, \dots, \Omega_N)$ , build the set of all  $d$ -dimensional embeddings of  $n$ -grams of the language,  $\mathfrak{N}_n(d), n = 1..N, d = 1 \dots D$ . Assumedly, the set of texts is a representative sample of all texts of the language under study.
- For a given  $n$ , use the sets of embeddings  $\mathfrak{N}_n(d), d = 1 \dots D$  construct the set of persistent homology classes (‘holes’) of the first order  $\{V_j\} = f(\{\mathfrak{N}_n(d)\})$  and ascertain their boundaries; one assumes that diameters  $d_j = \text{diam}V_j > d_{\max}$ , where  $d_{\max}$  is the maximum intracluster distance for clusters of synonyms of the language—this suggests that the holes thus found do belong to the large-scale

(coarse-grained) structure of the language and that they are not just local inhomogeneities of the sample.

For the bot-detection problem (to distinguish texts written by humans and those generated by bots), we use the following statement [6]: For a given natural language, one considers a space of texts  $\Omega$ , both written by humans and generated by bots. The space is divided into a subspace  $A = \{\alpha_1, \dots, \alpha_a\}$  of the texts written by humans and a subspace  $M = \bigcup_{j=1}^l M_j$  of texts generated by bots (ancient Greek  $\acute{\alpha}\nu\theta\rho\omega\pi\omicron\varsigma$ —a man, and  $\mu\eta\chi\acute{\alpha}\nu\eta\mu\alpha$ —a machine).  $M_j = \{\mu_1, \dots, \mu_{m_j}\}$  comprises text generated by the  $j$ th bot. The objective is to construct a set features  $\Lambda = \{\lambda_1, \dots, \lambda_k\}$  and to build a classifier  $R = R(\Lambda)$  with a classification error threshold  $r^*$ .

One randomly samples human-written texts in order to construct training and test sets. Most importantly, in order to construct training and test sets for bot-generated texts, one does not randomly sample this set of texts but randomly samples the set of bots themselves  $\{M_j, j = 1 \dots l\}$  into training and test subsets. The former comprises bots generating texts used to train the classifier; the latter comprises bots generating texts used to test it. The number of texts and the distribution of text sizes are approximately the same as those for the training and test sets for human-written texts.

### 1.3. The Algorithm to Reveal ‘Holes’ in a Natural Language

**1.3.1. The Semantic Space.** In order to obtain elements of the semantic space (embeddings of  $n$ -grams), we use the CBOW model [41]: This model places embeddings of words close in meaning near to each other; besides that, it ensures that the rules of ‘semantic arithmetic’ hold (for example, ‘woman’ + ‘king’ – ‘man’ = ‘queen’). The large-scale simulation ascertained the optimal dimensions of the semantic space (the number elements of embeddings),  $d = 100$  (we tested all possible values of  $d$  from 1 to 200). To construct embeddings of bigrams and trigrams, we concatenate the embeddings of the words it contains; thereby, the dimensions of the bigram and trigram semantic spaces are  $d = 100 * 2 = 200$ ,  $d = 100 * 3 = 300$ , respectively. We equip the space of word embeddings with the cosine distance (it is consistent with CBOW semantic arithmetic) and those of  $n$ -gram embeddings with the average minimum cosine distance over the pairs of words in  $n$ -grams:

$$\rho((a_1, a_2, \dots, a_n), (b_1, b_2, \dots, b_n)) = \frac{1}{n} \sum_{i \in \{1, \dots, n\}} \min_{j \in \{1, \dots, n\}} \text{cosine}(a_i, b_j). \quad (1)$$

For this metric, the distance between two permutations of  $n$ -grams (for instance, ‘to quickly walk’ and ‘to walk quickly’) is equal to zero.

**1.3.2. Persistent Homology.** To find homology ‘holes’ in the semantic space, homologies of the first order  $H_1$  (the number of homology classes is equal to the Betti number [42]; the homology classes of the zero order are the

connected components), we use the Vietoris–Rips filtration [42]:

**Definition 1.** The Vietoris–Rips complex of diameter  $\epsilon$  is the simplicial complex  $VR(\epsilon) = \{\sigma | \text{diam}(\sigma) \leq \epsilon\}$ , that is, a set of simplexes with diameters less than  $\epsilon$ .

**Definition 2.** The monotonically increasing sequence of real numbers  $\epsilon_1, \epsilon_2, \dots$  gives rise to the Vietoris–Rips filtration, that is, the sequence of the Vietoris–Rips complexes  $VR(\epsilon_1) \subseteq VR(\epsilon_2) \dots$

With the Vietoris–Rips filtration, one can determine the ‘birth’ and ‘death’ moments of various homology classes: The birth moment is defined as the value of  $\epsilon$  such that the new  $n$ -dimensional homology class emerges, due to the newly emerged  $n$ -dimensional simplex. The birth moment is defined as the value of  $\epsilon$  such that the new  $n$ -dimensional homology class disappears, due to the newly emerged  $(n + 1)$ -dimensional simplex.

#### 1.3.3. The Boundaries of ‘Holes’ in a Natural Language.

To explore the coarse-grained structure of a natural language, one should not only estimate the number of ‘holes’ (the Betti number) but also ascertain their boundaries, to find words/word-combinations that make the boundary of a hole and to contour (‘to drill out’) the hole. To this end, we employ the Čufar and Virk’s algorithm to reveal *homology representatives*, that is, elements one can employ to restore the respective homology class [43] (see also [44]).

The algorithm allows one to ascertain the chains of points that contour the holes (the boundaries of all holes). The core of the algorithm consists of two distinct phases: (1) to reduce the coboundary matrix and identify the simplices that contribute to the calculation of representative cycles and (2) to reduce the boundary matrix, focussing only on the columns identified in the first phase. Essentially, the reduced coboundary matrix is utilised to identify the death of simplices, disregarding all other columns during the boundary matrix reduction. Čufar and Virk formulated the following algorithm:

Input:  $X$ —finite metric space. Construct a fixed injective filtration function associated to the Rips filtration of  $X$ . Generate coboundary matrices  $d_k$ .

1. Reduce each coboundary matrix  $d_k$ . We may extract homological death simplices (i.e., cohomological birth simplices) and essential simplices.
2. For each  $k$  let  $D_k$  be the submatrix of the homology boundary matrix  $\partial_k$  consisting of columns corresponding to homological death simplices and essential simplices. We keep the indices of simplices to label the columns.
3. Compute persistent homology representatives by reducing  $D_k$  (The reduction algorithm is essentially a column reduction process from left to right. The core idea is to verify whether the boundary of the added simplex  $\sigma_i$  [i.e.,  $\text{Col}\partial(i)$ ] is homologically trivial

in  $K_{i-1}$  or not, by verifying whether it can be expressed as a linear combination of preceding columns).

**Output:** Return homology representatives from the reduced forms of matrices  $D_k$ .

By way of illustration, Figures 1 and 2 exhibit the revealed ‘holes’ (the homology classes of the first order) and their boundaries in two- and three-dimensional spaces. In the two-dimensional case, the algorithm reveals three holes (the diagram of Figure 1 demonstrated three persistent homologies), and in the three-dimensional case, two holes.

If one applies this algorithm to a natural language, one should take into account that, for a given hole, most points (embeddings) of the sample are not related to the hole in question—they are just located farther from the hole. To speed up the algorithm, it is reasonable to divide the semantic space into regions. To this end, we employ  $K$ -means clustering algorithms [45]. The algorithm minimises intercluster distances to obtain the clusters; it should be applied recurrently to obtain comparable clusters (five to ten thousand elements in each cluster). In order to single out the holes that belong to the large-scale (coarse-grained) structure of the language (not just caused by local inhomogeneities of the sample), we compare the holes’ diameters with those of the groups of synonyms for the language.

**1.3.4. The Bot Detection Problem.** To solve the bot detection problem, we, to construct classifiers, employ characteristics of the distributions of distances from the  $n$ -grams to the nearest holes. We deliberately use the simplest possible classifiers (support vector machines [SVM], decision trees [DT], and random forests [RF]) in order to distinguish texts of humans and those of bots. Importantly, to generate train and testing sets, we use different sets of bots, in order to ensure that the trained classifier would detect not only texts of the bots used to train it. Thus, out of the four LLM (mGPT, GPT-2, YaLM, and LSTM), we use two randomly selected ones to generate texts to train the classifier, and the two others to test it (there are six combinations to generate train and testing sets). To avoid the imbalanced testing set, we use an equal number of literature masterpieces and bot-generated texts (600 literature masterpieces and 300 texts of each LLMs used to generate the set of texts to train).

## 1.4. The Topological Structure of a Natural Language

**1.4.1. Data Collection.** To obtain word, bigram and trigram embeddings for the Russian and English languages, we employ the corpora of the national literature, downloadable from open sources. The total number of texts for the Russian language ( $|\mathfrak{S}_1|$ ) amounts to 6429 texts, with 103,952 unique words and 14775439 unique bigrams. The total number of texts for the English language ( $|\mathfrak{S}_2|$ ) amounts to 11,052 texts, with 94,087 unique words and 9,490,603 unique bigrams. To make the problem computationally tractable, we randomly select 10% and 1% of all words and bigrams (trigrams) available, respectively. In order to make the structure of the

bot texts comparable with the structure of that of a human, we prompt the LLMs (see above) to generate thousand-word long texts starting from the word from the corresponding texts of a human, thereby making the text of the same size as the corresponding human text, with each thousandth word coincided. We preprocess all texts (both natural and artificial) in the conventional way: the texts are tokenised and lemmatised; the proper nouns are replaced by the respective tokens (to this end, we use the Python libraries: *spacy* and *natasha* for the English and Russian languages, respectively).

**1.4.2. The General Structure of the Semantic Space.** In order to gain insight into the structure of the semantic space, we carried out a provisional experiment: we apply the above algorithm to the subsample (for  $d = 100$ ). To visualise data, we compress them into a two-dimensional space using the principle component analysis. The large-scale simulation allows us to conclude that the central area of the point cloud is more densely populated than ‘the outskirts’. Consequently, the holes around the centre are smaller and die earlier (cf. Figures 3, 4, and 5). Figure 6 shows the respective distributions (in the central region, lifetime amounts to 0.15; whereas in the outskirts, 0.24). Finally, Figure 7 illustrates the ratio of homology classes’ persistence and hole diameters. Evidently, the values for the holes in the central region of the space of words are smaller. In what follows, we study the holes in the central region and in the outskirts separately, and use different threshold values for them.

### 1.4.3. First Order Homology in the Semantic Space

**1.4.3.1. The Boundaries.** The algorithm discussed above is applied to delineate homologies of the zero and first order in the semantic spaces of words, bigrams and trigrams for the English and Russian languages: Figures 8 and 9 show the respective persistence diagrams; Table 1 summarises the total number of the homology classes. Clearly, most homology classes are not persistent and die early—they produce, in the persistent diagrams, points close to the birth = death diagonal (Figures 8 and 9 for the English and Russian languages, respectively).

Interestingly, for the English and Russian languages, the persistent diagrams resemble one another: The homology classes are less persistent for the words (the orange strip is narrower; points reside closer to the diagonal, which implies that homology classes die instantly) than for the bigrams and trigrams (there is a number of points far from the diagonal). One can conclude that the space of bigrams and trigrams ‘betray’ the immanent semantic structure of the respective language.

**1.4.3.2. The Most Persistent Homologies.** To obtain the most persistent homology classes, we construct the distributions of the homology diameters in order to use those of 0.99 quantile as candidates: Figures 10 and 11 exhibit the distributions for the English and Russian languages, respectively.

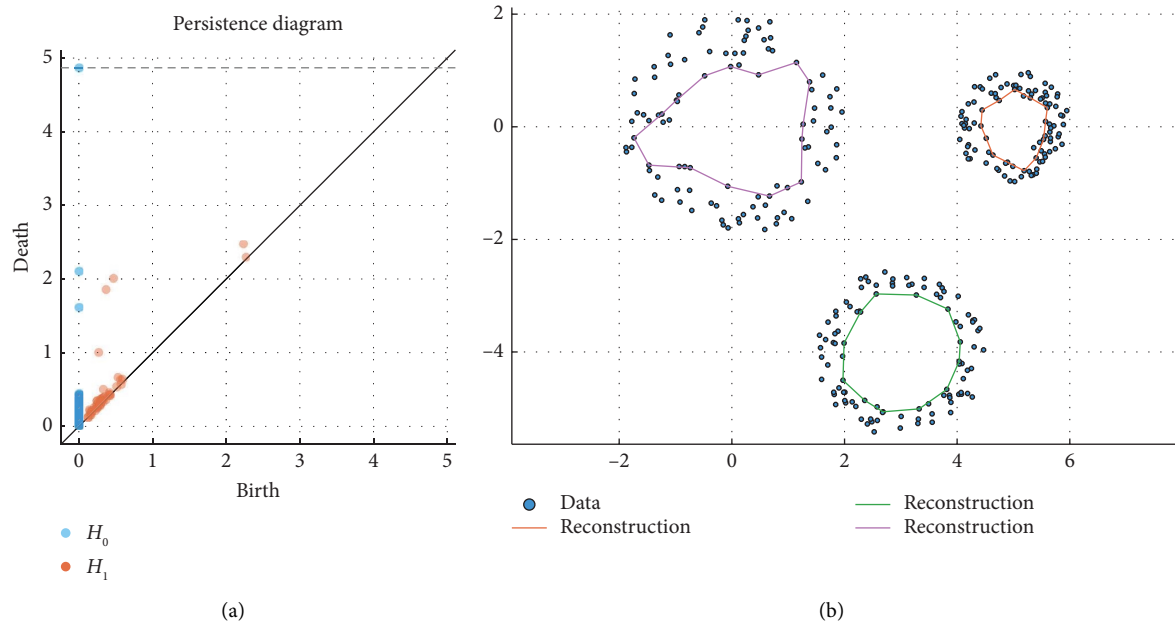


FIGURE 1: The persistence diagram (a) and boundaries of the first-order homology classes (b) for synthetic data in the two-dimensional space.

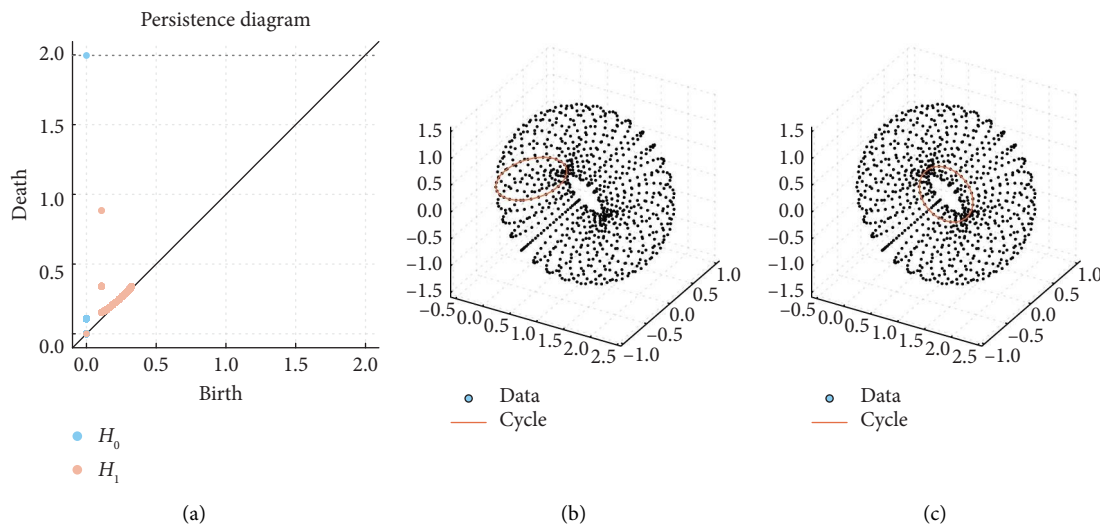


FIGURE 2: The persistence diagram (a) and boundaries of two first-order homology classes (b, c) for synthetic data (a two-dimensional torus) in the three-dimensional space.

As the words ( $n$ -grams) close in meaning reside closely in the semantic space, a hole of small diameters comprises just synonyms; thereby, they are not good candidates for ‘the blind spots’ of the language. Respectively, to single out holes that do belong to the coarse-grained structure of the language, we compare their diameters with the maximum diameter over the groups of synonyms for the respective languages (for this purpose, we use the vocabularies of synonyms for the Russian (“The Vocabulary of Synonyms of the Russian Language” [Ed. L. Babenko, 2011]) and English languages (the model *wordnet* of the library *nlk*); an

example of a synonym group: ‘bending, bend, twist, bow, bight, etc. For  $n$ -grams, we use all combinations of all groups of synonyms (for example, for the bigram ‘unstable shore’, we take all combinations of the next groups of synonyms ‘fragile, shaky, unsteady, unstable and wobble’ and ‘bank, shore, coast, beach, side and coastline’) and calculate the maximum diameter over all groups of such combinations to use it as the threshold to single out the holes in the space of  $n$ -grams (refer to the distance function (1)) (the right-most vertical lines in Figures 10 and 11). This yields the following number of holes:

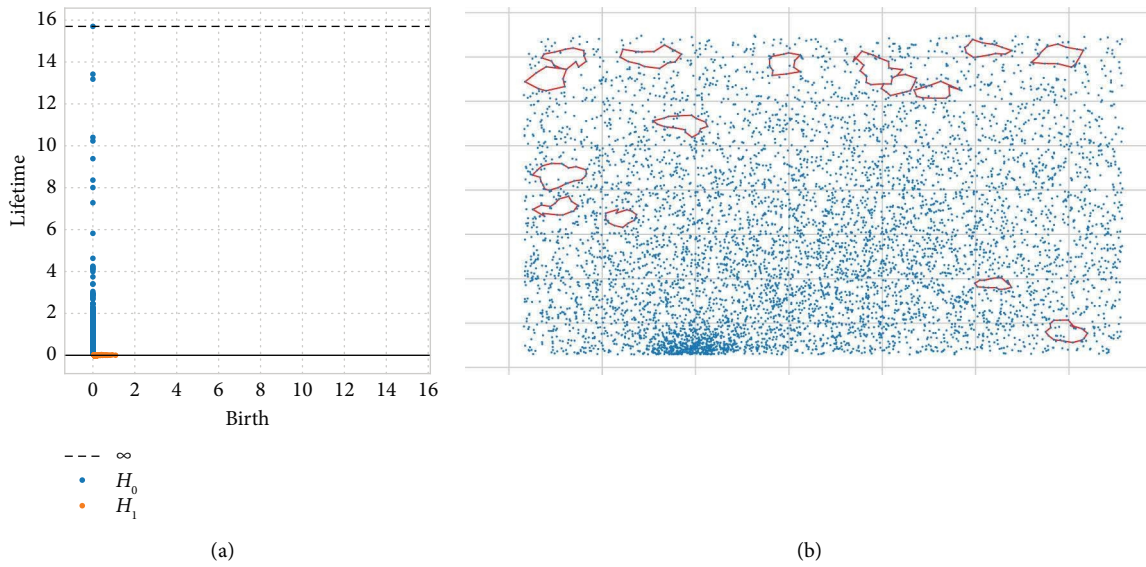


FIGURE 3: Homologies in the language central region (two-dimensional PCA projection): (a) persistence diagrams for homologies of the zero and first order; (b) boundaries of the most persistent diagrams.

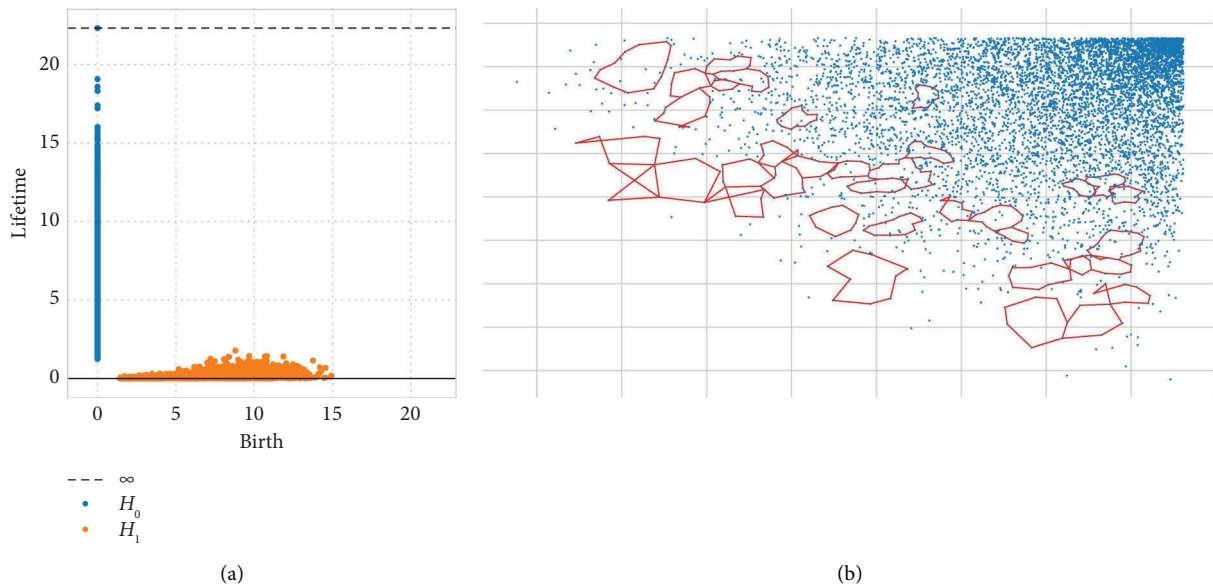


FIGURE 4: Homologies in the language outskirts (two-dimensional PCA projection): (a) persistence diagrams for homologies of the zero and first order; (b) boundaries of the most persistent diagrams.

- Seven holes in the space of words of the Russian language;
- Twelve holes in the space of bigrams of the Russian language;
- Twelve holes in the space of trigrams of the Russian language;
- Four holes in the space of words of the English language;
- Five holes in the space of bigrams of the English language;

- Seven holes in the space of trigrams of the English language.

An example of the boundaries are as follows.

- giggly, dark-eyed, dark-brown, [young married woman], to expatiate, there's, so-and-so, to put in shame, to scold, [to be angry], [to be confused], to confuse, confused, embarrassed, puzzled, to dumb-found, [to be struck dumb], baffled by, pointedly, well-disposed, kindly, friendly, affable, good-natured, ingenuous, sprightly, smart;

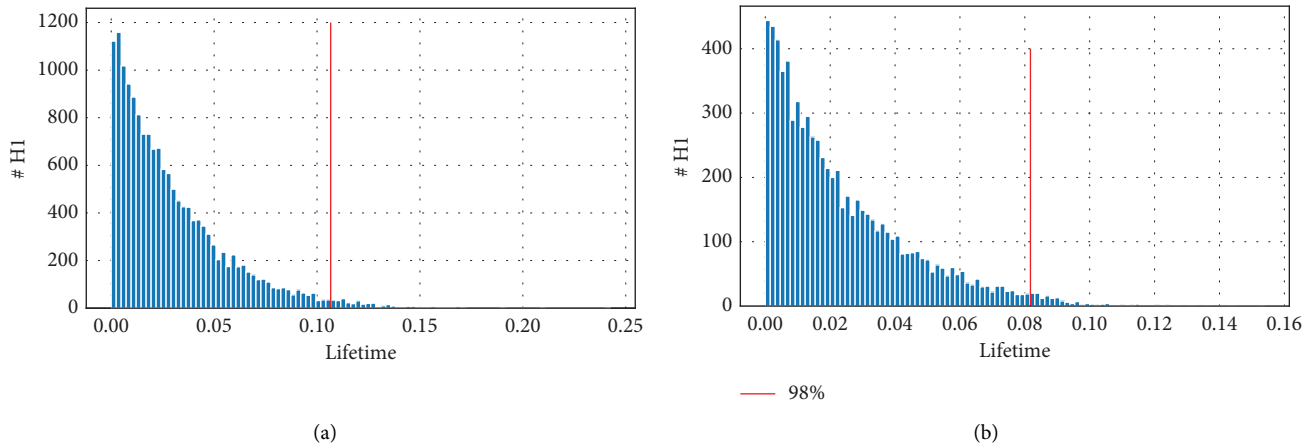


FIGURE 5: The distribution of the first-order homology persistence (a) in the outskirts and (b) in the central region. The vertical red lines denote 96% quantiles.

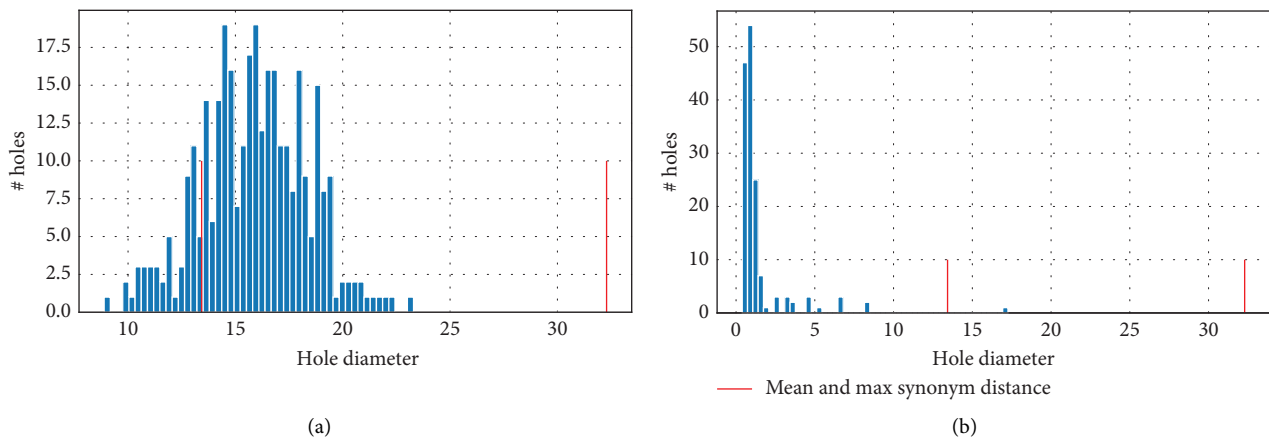


FIGURE 6: The distribution of the first-order homology diameters (a) in the outskirts and (b) in the central region. The vertical red lines denote the average and maximum diameters of the group of synonyms.

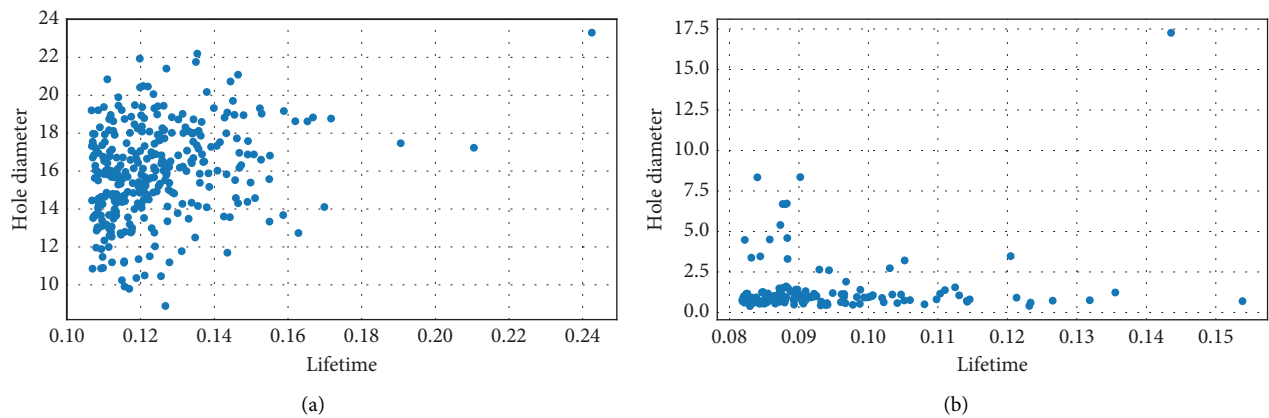


FIGURE 7: The ratio of persistence and diameters for the first-order homology classes (a) in the outskirts and (b) in the central region for the space of words (the Russian language).

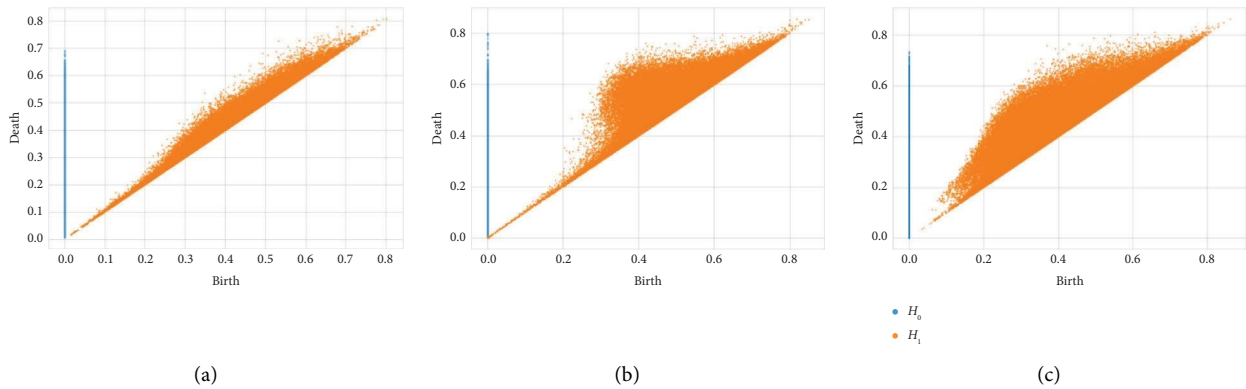


FIGURE 8: The persistent diagrams for (a) words, (b) bigrams, and (c) trigrams for the English language. The blue points correspond to the homology classes of the zero order, and the orange ones to the homology classes of the first order.

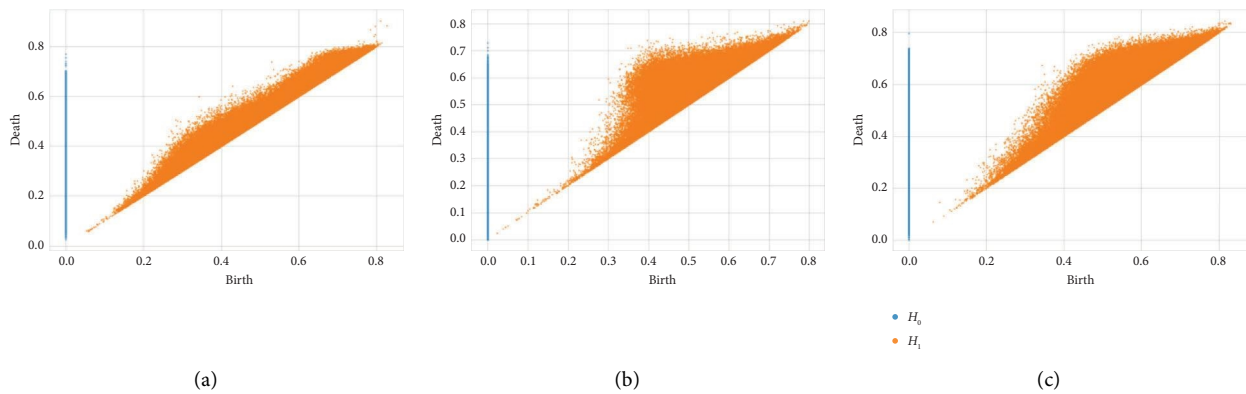


FIGURE 9: The persistent diagrams for (a) words, (b) bigrams and (c) trigrams for the Russian language. The blue points correspond to the homology classes of the zero order, and the orange ones to the homology classes of the first order.

TABLE 1: The total number of homology classes of the first order.

	Words	Bigrams	Trigrams
Russian	153,813	225,488	227,724
English	77,477	127,436	110,472

- [go well] principality, [Peter the Great] empire, Chinese state, German power, Latvian authority, speak power, [so much] might, [so much] cordiality, [so much] love, jealousy love, to look love, to look life, to look or, [naughty child] or, emotional or, emotional future, [run by] future, many future, many ancestor, omniscient descendant;
- home cornfield own, home town attack, home commander-in-chief frontline, name commander-in-chief frontline, former officer officer, together regiment mate, own troop mate, own orderly Orenburg, own cornfield rely, own Lyon [fellow countryman].

It is worthy to note that these chains (boundaries) can contain synonyms, but they do not contain only synonyms; the ‘linear combination’ of nonsynonymic elements of such

a boundary gives meanings that are unavailable in the respective language, thereby betraying its holes, blind spots. The lists of words/bigrams/trigrams of the boundaries are available at [https://github.com/quynhu-d/stb-tda/tree/main/hole\\_contours](https://github.com/quynhu-d/stb-tda/tree/main/hole_contours).

1.4.3.3. *Texts and Holes.* For the bot-detection problem, it is reasonable to calculate the following quantities (both for human-written and bot-generated texts; the quantity is averaged over words, bigrams or trigrams of the text in question):

- The distance to the centre of each homology class;
- The minimum distance to each homology class;
- The maximum distance to each homology class;
- The minimum distance to the nearest homology class.

Figures 12, 13 and 14 exhibit the distributions of the minimum and maximum distances to homology classes (averaged over the texts of the sample), respectively, for the Russian language; Figure 12 (the distribution for words) betrays the fact that the distribution for human texts (blue colour) is statistically significantly shifted with respect to that for the text of bots (orange). The same holds true for the

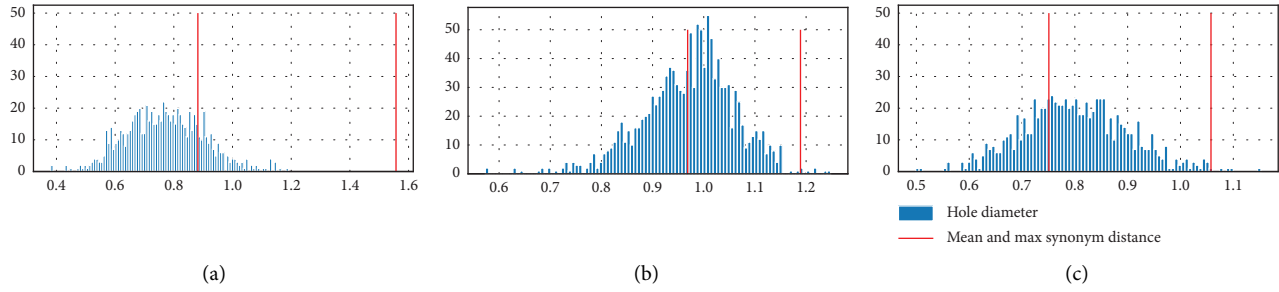


FIGURE 10: The distribution of diameters for the first-order homology classes for (a) words, (b) bigrams and (c) trigrams for the English language. The red vertical lines denote the average and maximum diameters of the groups of synonyms.

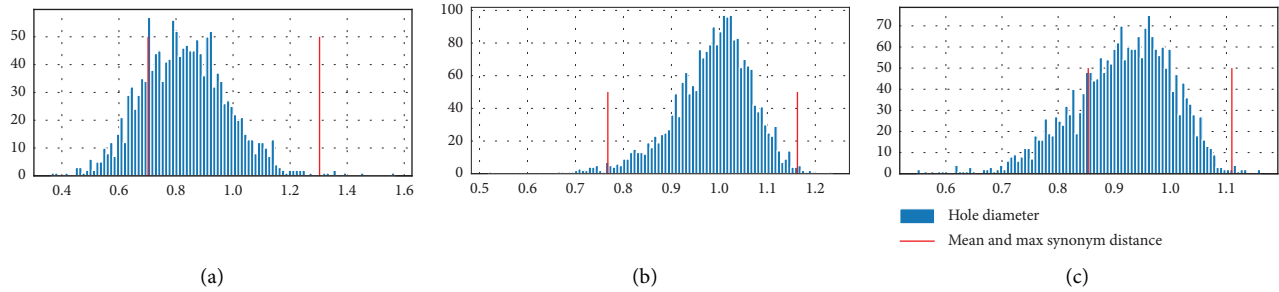


FIGURE 11: The distribution of diameters for the first-order homology classes for (a) words, (b) bigrams and (c) trigrams for the Russian language. The red vertical lines denote the average and maximum diameters of the groups of synonyms.

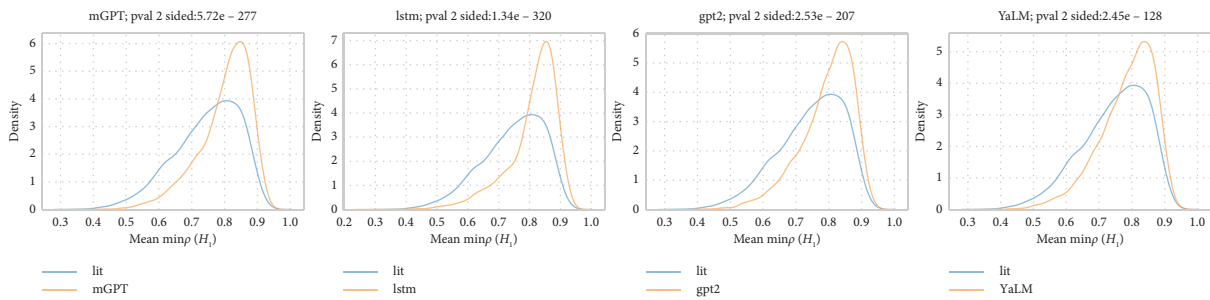


FIGURE 12: The distribution of the averaged minimum distance to the homology classes for the space of words (the Russian language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from right to the left: mGPT, LSTM, GPT-2 and YaLM).

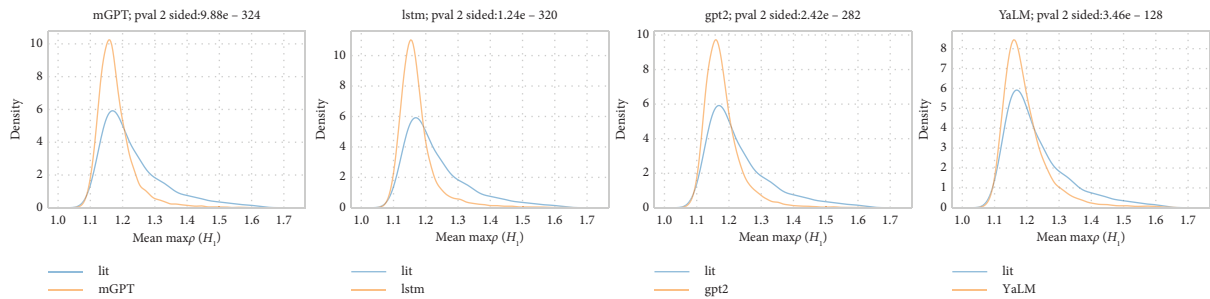


FIGURE 13: The distribution of the averaged maximum distance to the homology classes for the space of words (the Russian language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from the right to the left: mGPT, LSTM, GPT-2 and YaLM).

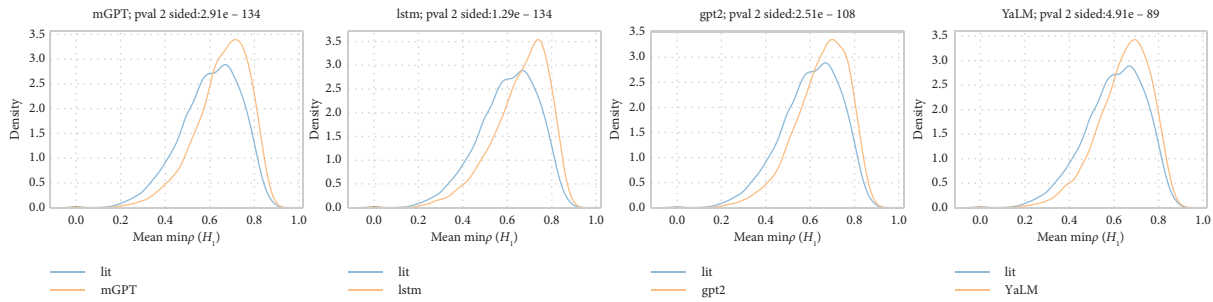


FIGURE 14: The distribution of the averaged minimum distance to the nearest homology classes for the space of words (the Russian language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from the right to the left: mGPT, LSTM, GPT-2 and YaLM).

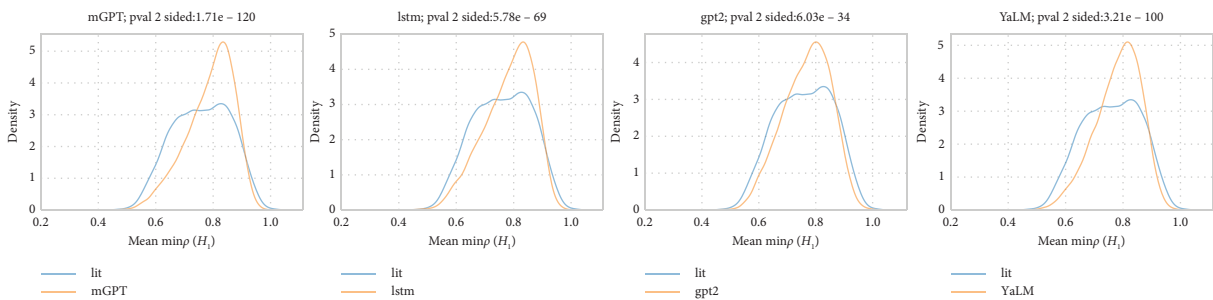


FIGURE 15: The distribution of the averaged minimum distance to the homology classes for the space of words (the English language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from right to the left: mGPT, LSTM, GPT-2 and YaLM).

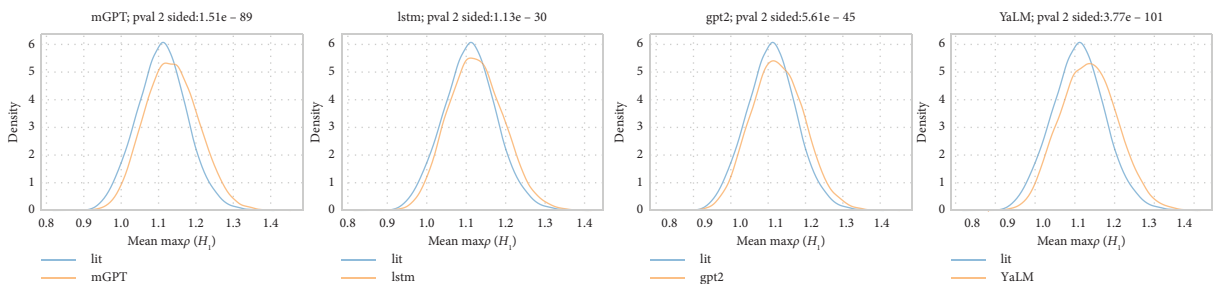


FIGURE 16: The distribution of the averaged maximum distance to the homology classes for the space of words (the English language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from the right to the left: mGPT, LSTM, GPT-2 and YaLM).

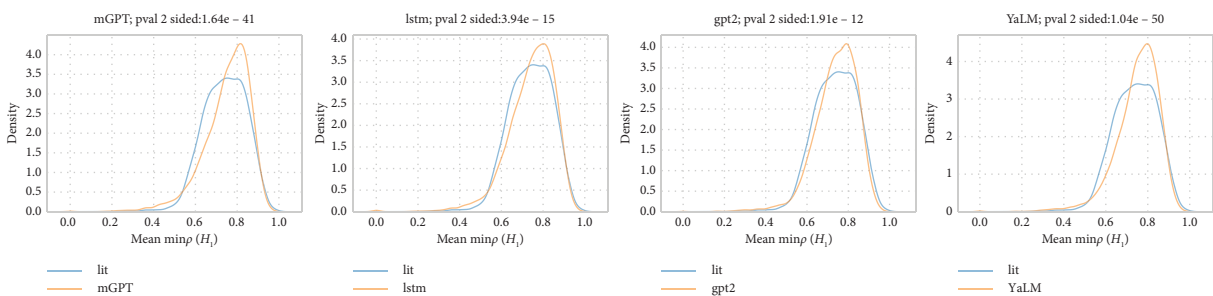


FIGURE 17: The distribution of the averaged minimum distance to the nearest homology classes for the space of words (the English language). The blue curve corresponds to the literary text and the orange one to the bot-generated texts (from the right to the left: mGPT, LSTM, GPT-2 and YaLM).

TABLE 2: Classification accuracy for the English language.

	Train $M_1, M_2$	Test $M_3, M_4$	Train $M_1, M_3$	Test $M_2, M_4$	Train $M_1, M_4$	Test $M_2, M_3$	Train $M_2, M_3$	Test $M_1, M_4$	Train $M_2, M_4$	Test $M_1, M_3$	Train $M_3, M_4$	Test $M_1, M_2$
<i>Words</i>												
SVM	0.879	0.922	0.919	0.903	0.878	0.771	0.96	0.844	0.958	0.84	0.972	0.803
DT	0.99	0.932	0.983	0.889	0.973	0.768	0.983	0.85	0.949	0.828	0.986	0.84
RF	0.998	<b>0.935</b>	0.983	<b>0.92</b>	0.98	<b>0.773</b>	0.998	<b>0.857</b>	0.989	<b>0.861</b>	0.997	<b>0.852</b>
<i>Bigrams</i>												
SVM	0.947	0.94	0.972	0.925	0.964	<b>0.868</b>	0.963	<b>0.941</b>	0.959	0.94	0.968	0.878
DT	0.975	0.935	1	0.914	0.989	0.866	0.977	0.879	1	0.918	0.993	0.866
RF	0.99	<b>0.964</b>	0.999	<b>0.929</b>	0.997	0.856	0.99	0.93	1	<b>0.963</b>	0.996	<b>0.896</b>
<i>Trigrams</i>												
SVM	0.922	0.886	0.952	0.83	0.968	<b>0.841</b>	0.906	0.755	0.893	0.764	0.935	0.722
DT	0.969	0.878	1	0.813	0.941	0.793	0.957	0.819	0.976	0.902	0.982	<b>0.884</b>
RF	0.976	<b>0.914</b>	1	<b>0.883</b>	0.994	0.747	0.993	<b>0.928</b>	1	<b>0.909</b>	0.993	0.87

Note:  $M_1$  stands for LSTM;  $M_2$ , for YaLM;  $M_3$ , for GPT-2;  $M_4$ , for mGPT. Random forest outperforms other models in most settings, distinguishing one set from another with the highest accuracy. SVM yields better classification accuracy for YaLM and GPT-2 (bigrams and trigrams), as well as LSTM and mGPT (bigrams). Bold values represent best metrics (among the three models, decision tree, random forest and support vector machine) for each test set. Abbreviations: DT, decision tree; RF, random forest; SVM, support vector machine.

TABLE 3: Classification accuracy for the Russian language.

	Train $M_1, M_2$	Test $M_3, M_4$	Train $M_1, M_3$	Test $M_2, M_4$	Train $M_1, M_4$	Test $M_2, M_3$	Train $M_2, M_3$	Test $M_1, M_4$	Train $M_2, M_4$	Test $M_1, M_3$	Train $M_3, M_4$	Test $M_1, M_2$
<i>Words</i>												
SVM	0.93	<b>0.916</b>	0.947	0.895	0.939	0.952	0.983	0.743	0.963	0.779	0.971	0.729
DT	0.979	0.908	0.932	<b>0.926</b>	0.98	0.928	0.994	<b>0.774</b>	0.982	<b>0.868</b>	0.964	<b>0.779</b>
RF	0.998	0.895	1	0.921	0.974	<b>0.945</b>	0.998	0.756	1	0.795	0.992	0.774
<i>Bigrams</i>												
SVM	0.962	<b>0.937</b>	0.975	0.852	0.986	0.79	0.972	0.714	0.974	0.727	0.988	0.703
DT	0.995	0.886	0.922	0.876	0.981	0.828	0.99	0.705	1	0.712	1	0.708
RF	1	0.929	1	<b>0.917</b>	0.995	<b>0.911</b>	1	<b>0.726</b>	1	<b>0.737</b>	0.999	<b>0.712</b>
<i>Trigrams</i>												
SVM	0.963	0.934	0.971	0.924	0.964	0.885	0.961	0.718	0.965	0.721	0.961	0.705
DT	0.987	0.886	0.98	0.883	0.987	0.859	0.989	<b>0.822</b>	0.992	<b>0.927</b>	0.992	<b>0.737</b>
RF	0.993	<b>0.943</b>	1	<b>0.947</b>	0.997	<b>0.879</b>	1	0.755	1	0.737	1	0.735

Note:  $M_1$  stands for LSTM;  $M_2$ , for YaLM;  $M_3$ , for GPT-2;  $M_4$ , for mGPT. A decision tree model demonstrates superior performance on the majority of test sets for word-level features. For bigrams, the random forest model achieves the highest accuracy in five out of six test sets. At the trigram level, both random forest and decision tree models exhibit comparable, leading performance. Bold values represent best metrics for each test set. Abbreviations: DT, decision tree; RF, random forest; SVM, support vector machine.

distributions for the nearest homology classes (Figure 14). Figures 15, 16 and 17 show similar distributions for the English language. In general, the words and  $n$ -grams of the texts of bots reside statistically farther from the boundaries of homology classes (the boundaries of the language) than those of the texts of humans. Formally, we put forward the null hypothesis  $H_0$  that the distance between a text and the nearest ‘hole’ (no matter how we measure it) differs for humans and bots and test it by the Kolmogorov–Smirnov (nonparametric) criterion [46]. For all the above characteristics, the  $p$  value appears to be less than 0.05; this allows one to reject the null hypothesis that samples are generated by the same distribution (interestingly, the number of the nearest hole also differs for human and bot texts. For instance, for the Russian language, for the most human texts, the nearest hole for words is Hole #5 [see [\[word\\\_hole\\\_contours.txt\]\(https://github.com/quynhu-d/stb-tda/blob/main/hole\_contours.txt\) for details\], whereas, for most bot texts, the nearest hole for words is Hole #1 \[40% words of mGPT-texts; 44% words of LSTM; 38% words of GPT-2-texts; 32% words of YaLM-texts\], whereas only 8%–10% of words are close to Hole #5\).](https://github.com/quynhu-d/stb-tda/blob/main/hole_contours/RU/ru_</a></p>
</div>
<div data-bbox=)

Consequently, one can use the following characteristics to solve the bot-detection problem:

1. The distance from a text to the centre of each homology class (averaged distance of words/bigrams/trigrams to the centre of a homology class);
2. The average distance from a text to the centres of homology classes (averaged distances of words/bigrams/trigrams);
3. The minimum distance from a text to each homology class (averaged minimum distances of words/bigrams/trigrams to a homology class);

4. The average minimum distance from a text to homology classes (averaged minimum distances of words/bigrams/trigrams);
5. The maximum distance from a text to each homology class (averaged maximum distances of words/bigrams/trigrams to a homology class);
6. The average maximum distance from a text to homology classes (averaged maximum distances of words/bigrams/trigrams);
7. The share of words/bigrams/trigrams closest to a given homology class.

For each text  $(n\_holes + 1) * 4$  characteristics are calculated, where  $n\_holes$  is the number of found holes (see previous section).

It is worthy to note that quantities analysed in [47] can also be used as such characteristics.

**1.5. Classifiers: Humans and Bots.** We deliberately use the simplest possible classifier models, in order to assess the characteristics employed. We use different texts generated by different set of bots to generate training and test sets (see *Problem statement*). To find optimal hyperparameters for the trained models used, we employ 10-fold cross-validation; for the SVM, to find the optimal parameter  $C$ , we search over the range from  $1e-5$  to 10; for the DT, to find optimal depth and number of elements in the leaves, we search over the ranges from 3 to 15 and from 1 to 4, respectively.

Tables 2 and 3 summarise classification accuracy for the English and Russian languages, respectively (the samples are balanced). Despite the fact that we train and test classifiers on texts generated by different set of bots, the performance of all models is greater than 0.71. Accuracy of the best model, averaged over all possible combinations of training and test bots, is more than 0.8: For the English language, it amounts to 0.87 for words, 0.93 for bigrams and 0.89 for trigrams; for the Russian language, it amounts to 0.86 for words, 0.82 for bigrams and 0.88 for trigrams. In the English language, the best classifier, in most cases, appears to be the RF.

## 2. Conclusions

The present paper, in the frameworks of scientific methodology, analyses the most fundamental, topological properties of *Hailanokea* (the set of language fractal structures of embeddings of all words, bigrams and trigrams). To this end, it employs TDA methods to reveal ‘holes’ (stable persistent homology classes) in the semantic spaces of words, bigrams and trigrams of the English and Russian languages and to ascertain their boundaries. Furthermore, the paper selects those holes that belong to the large-scale (coarse-grained) structure of the language that are not just local inhomogeneities of the sample—it appears that there are around a dozen of them for each of the languages (English and Russian). These boundaries delineate ‘blind spots’ of the respective language (the regions of the semantic spaces that do not contain words/bigrams/trigrams of the language—that is, regions of concepts that the language cannot see through its lens).

The secondary goal of the paper is to solve the bot-detection problem in its strong statement, that is, one trains the classifiers on one set of bots and tests on another set of bots. To this end, we estimate the average distances from words, bigrams and trigrams of a text to the boundaries of the nearest ‘hole’, for texts both written by humans and generated by bots, and construct classifiers. The classifiers show comparatively good results: the average accuracy is more than 0.8.

The future research direction implies exploring words/bigrams/trigrams of the boundaries of the language holes—it is possible to match the revealed homology classes to meaningful semantic or formal linguistic structures. The present paper explores the homology of the first order (topological ‘holes’); it would be interesting to explore homologies of the higher orders, in particular, those of the second order (‘voids’). We also expect to investigate languages of other language families and those with distinctly different values of parameters of the universal grammar.

## Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## Disclosure

An earlier version of this article is available as a preprint [48].

## Conflicts of Interest

The authors declare no conflicts of interest.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## Acknowledgements

The authors are sincerely indebted to Ms. A. Kogan, HSE for natural language texts preprocessing and embedding construction for the Russian and English languages.

The authors are indebted to Mr. J. Cumberland, HSE for text-editing and proof-reading.

This research was supported in part through computational resources of HPC facilities at HSE University.

## References

- [1] V. A. Gromov and A. M. Migrina, “A Language as a Self-Organized Critical System,” *Complexity* 2017, no. 1 (2017): 9212538.
- [2] J. Hellrich, *Word Embeddings: Reliability & Semantic Change* (IOS Press, 2019).
- [3] V. A. Gromov, N. S. Borodin, and A. S. Yerbolova, “A Language and Its Dimensions: Intrinsic Dimensions of Language Fractal Structures,” *Complexity* 2024, no. 1 (2024): 8863360, <https://doi.org/10.1155/2024/8863360>.

- [4] W. Harbert, *The Germanic Languages* (Cambridge University Press, 2006).
- [5] M. S. Dryer and T. Shopen, *Language Typology and Syntactic Description* (2007).
- [6] V. A. Gromov, Q. N. Dang, A. S. Kogan, and A. S. Yerbolova, “Spot the Bot: the Inverse Problems of NLP,” *Peer Journal of Computer Science*.
- [7] N. Allott, T. Lohndal, and G. Rey, “Synoptic Introduction,” *A Companion to Chomsky* (2021), 1–17.
- [8] Ł. Dębowski, *Information Theory Meets Power Laws: Stochastic Processes and Language Models* (John Wiley & Sons, 2020).
- [9] K. Tanaka-Ishii, “Language as a Complex System,” *Statistical Universals of Language: Mathematical Chance vs. Human Choice* (Springer Nature, 2021).
- [10] S. Semple, R. Ferrer-i-Cancho, and M. L. Gustison, “Linguistic Laws in Biology,” *Trends in Ecology & Evolution* 37, no. 1 (2022): 53–66, <https://doi.org/10.1016/j.tree.2021.08.012>.
- [11] V. A. Gromov and Q. N. Dang, “Semantic and Sentiment Trajectories of Literary Masterpieces,” *Chaos, Solitons & Fractals* 175 (2023): 113934, <https://doi.org/10.1016/j.chaos.2023.113934>.
- [12] X. Wu, J. Yu, and X. Zhao, “Spatio-Temporal Keyword Query in Semantic Trajectories,” *Frontiers of Computer Science* 16, no. 2 (2022): 162602, <https://doi.org/10.1007/s11704-020-0039-4>.
- [13] V. A. Gromov, et al., “Comparative Study of Natural Languages as Self-Organised Critical Systems,” *Chaos, Solitons & Fractals*.
- [14] M. Garg, A. K. Gupta, and R. Prasad, *Graph Learning and Network Science for Natural Language Processing* (CRC Press, 2022).
- [15] M. Garg and M. Kumar, “The Structure of Word Co-Occurrence Network for Microblogs,” *Physica A: Statistical Mechanics and Its Applications* 512 (2018): 698–720, <https://doi.org/10.1016/j.physa.2018.08.002>.
- [16] T. Stanisz, S. Drożdż, and J. Kwapien, “Complex Systems Approach to Natural Language,” *Physics Reports* 1053 (2024): 1–84, <https://doi.org/10.1016/j.physrep.2023.12.002>.
- [17] M. E. Aktas, E. Akbas, and El F. Ahmed, “Persistence Homology of Networks: Methods and Applications,” *Applied Network Science* 4, no. 1 (2019): 1–28.
- [18] D. Horak, S. Maletić, and M. Rajković, “Persistent Homology of Complex Networks,” *Journal of Statistical Mechanics: Theory and Experiment* 2009, no. 03 (2009): P03034, <https://doi.org/10.1088/1742-5468/2009/03/p03034>.
- [19] A. Myers, E. Munch, and F. A. Khasawneh, “Persistent Homology of Complex Networks for Dynamic State Detection,” *Physical Review E* 100, no. 2 (2019): 022314, <https://doi.org/10.1103/physreve.100.022314>.
- [20] X. Xu, J. Cisewski-Kehe, S. B. Green, and D. Nagai, “Finding Cosmic Voids and Filament Loops Using Topological Data Analysis,” *Astronomy and Computing* 27 (2019): 34–52, <https://doi.org/10.1016/j.ascom.2019.02.003>.
- [21] J. Beuria, “Persistent Homology of Collider Observations: When (W) Hole Matters,” *Physics Letters B* 846 (2023): 138188, <https://doi.org/10.1016/j.physletb.2023.138188>.
- [22] R. Bermejo, G. Wilding, R. van de Weygaert, B. J. T. Jones, G. Vegter, and K. Efstathiou, “Topological Bias: How Haloes Trace Structural Patterns in the Cosmic Web,” *Monthly Notices of the Royal Astronomical Society* 529, no. 4 (2024): 4325–4353, <https://doi.org/10.1093/mnras/stae543>.
- [23] Y. Skaf and R. Laubenbacher, “Topological Data Analysis in Biomedicine: A Review,” *Journal of Biomedical Informatics* 130 (2022): 104082, <https://doi.org/10.1016/j.jbi.2022.104082>.
- [24] Z. Meng, D. V. Anand, Y. Lu, J. Wu, and K. Xia, “Weighted Persistent Homology for Biomolecular Data Analysis,” *Scientific Reports* 10, no. 1 (2020): 2079, <https://doi.org/10.1038/s41598-019-55660-3>.
- [25] T. K. Dey and S. Mandal, “Protein Classification With Improved Topological Data Analysis,” in *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)* (Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2018).
- [26] P. Corcoran and C. B. Jones, “Topological Data Analysis for Geographical Information Science Using Persistent Homology,” *International Journal of Geographical Information Science* 37, no. 3 (2023): 712–745, <https://doi.org/10.1080/13658816.2022.2155654>.
- [27] L. Caputi, A. Pidnebesna, and J. Hlinka, “Promises and Pitfalls of Topological Data Analysis for Brain Connectivity Analysis,” *NeuroImage* 238 (2021): 118245, <https://doi.org/10.1016/j.neuroimage.2021.118245>.
- [28] J. Yoo, E. Y. Kim, Y. M. Ahn, and J. C. Ye, “Topological Persistence Vineyard for Dynamic Functional Brain Connectivity During Resting and Gaming Stages,” *Journal of Neuroscience Methods* 267 (2016): 1–13, <https://doi.org/10.1016/j.jneumeth.2016.04.001>.
- [29] H. Walter, “Topologies on Formal Languages,” *Mathematical Systems Theory* 9, no. 2 (1975): 142–158, <https://doi.org/10.1007/bf01704017>.
- [30] S. Y. Kuroda, “A Topological Study of Context-Free Languages,” *Journées d’Etudes sur Analyse Syntactique* (1969).
- [31] S. A. Fulop and D. Kephart, “Topology of Language Classes,” in *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)* (2015), 26–38, <https://doi.org/10.3115/v1/w15-2303>.
- [32] L. Qi, Y. Wang, J. Chen, M. Liao, and J. Zhang, “Culture Under Complex Perspective: a Classification for Traditional Chinese Cultural Elements Based on Nlp and Complex Networks,” *Complexity* 2021, no. 1 (2021): 6693753, <https://doi.org/10.1155/2021/6693753>.
- [33] B. Benigni, M. Dallabona, E. Bravi, S. Merler, and M. De Domenico, “Navigating Concepts in the Human Mind Unravels the Latent Geometry of Its Semantic Space,” *Complexity* 2021, no. 1 (2021): 6398407, <https://doi.org/10.1155/2021/6398407>.
- [34] X. Zhu, “Persistent Homology: An Introduction and a New Text Representation for Natural Language Processing,” in *IJCAI* (2013), 1953–1959.
- [35] N. Elyasi and M. Hosseini Moghadam, “An Introduction to a New Text Classification and Visualization for Natural Language Processing Using Topological Data Analysis” (2019).
- [36] K. Savle, W. Zadrozny, and M. Lee, “Topological Data Analysis for Discourse Semantics?” in *Proceedings of the 13th International Conference on Computational Semantics-Student Papers* (2019), 34–43, <https://doi.org/10.18653/v1/w19-0605>.
- [37] S. Tymochko, J. Chaput, T. Doster, E. Purvine, W. Jackson, and T. Emerson, “Con Connections: Detecting Fraud from Abstracts Using Topological Data Analysis,” in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (IEEE, 2021), 403–408.
- [38] A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta, “Topology of the Conceptual Network of Language,” *Physical*

- Review E* 65, no. 6 (2002): 065102, <https://doi.org/10.1103/physreve.65.065102>.
- [39] S. Fitz, “The Shape of Words-Topological Structure in Natural Language Data,” in *Topological, Algebraic and Geometric Learning Workshops 2022* (PMLR, 2022), 116–123.
- [40] A. Rathore, “Topological Data Analysis and Visualization for Interpretable Machine Learning,” (The University of Utah, 2023), PhD dissertation.
- [41] T. Mikolov, “Efficient Estimation of Word Representations in Vector Space” (2013).
- [42] H. Edelsbrunner and J. L. Harer, *Computational Topology: An Introduction* (American Mathematical Society, 2022).
- [43] M. Čufar and Ž. Virk, “Fast Computation of Persistent Homology Representatives With Involved Persistent Homology,” *arXiv preprint arXiv:2105.03629* (2021).
- [44] I. Obayashi, “Stable Volumes for Persistent Homology,” *Journal of Applied and Computational Topology* 7, no. 4 (2023): 671–706, <https://doi.org/10.1007/s41468-023-00119-8>.
- [45] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” in *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (University of California Press, 1967).
- [46] N. V. Smirnov, “[Approximation of Distribution Laws of Random Variables Based on Empirical Data] Priblizhenie Zakonov Raspredelenia Sluchainykh Velichin po Empiricheskim Dannym,” *Uspekhi Matematicheskikh Nauk* 10 (1944): 179–206.
- [47] S. Fitz, P. Romero, and J. J. Schneider, “Hidden Holes: Topological Aspects of Language Models” (2024).
- [48] V. A. Gromov, Q. N. Dang, and A. Yerbolova, “A Language and Its Holes: The First Order Homologies of the Large-Scale Geometrical Structure of a Natural Language,” *ResearchGate Preprint* (2024): [https://www.researchgate.net/publication/384012422\\_A\\_Language\\_and\\_Its\\_Holes\\_the\\_First\\_Order\\_Homologies\\_of\\_the\\_Large-scale\\_Geometrical\\_Structure\\_of\\_a\\_Natural\\_Language](https://www.researchgate.net/publication/384012422_A_Language_and_Its_Holes_the_First_Order_Homologies_of_the_Large-scale_Geometrical_Structure_of_a_Natural_Language).