UDC 004.912:004.021

DOI: 10.15587/1729-4061.2026.351509

REVEALING INTRINSIC DIMENSIONALITY PATTERNS IN SEMANTIC SPACES OF NATURAL LANGUAGES USING GRAPH ALGORITHMS

Assel Yerbolova
Corresponding author
Master of Computer Sciences, Doctoral Student*
E-mail: erbolova1983@mail.ru
ORCID: https://orcid.org/0009-0007-7119-4665

Ildar Kurmashev
Candidate of Technical Sciences, Associate Professor*
ORCID: https://orcid.org/0000-0001-9872-7483

*Department of Information
and Communication Technologies
Manash Kozybayev North Kazakhstan University
Pushkin str., 86, Petropavlovsk,
Republic of Kazakhstan, 150000

This study considers semantic spaces of n-grams (unigrams, bigrams, and trigrams) formed from natural language texts. The problem under consideration is related to the limitations of conventional approaches, which use semantic spaces of a fixed high dimensionality without taking into account their internal geometric structure. An experimental study of the internal dimensionality of vector representations of linguistic objects used in natural language processing tasks was conducted.

To solve the set task, graph algorithms for estimating internal dimension were applied. These algorithms are based on the analysis of minimum spanning tree statistics, allowing for estimates of both Hausdorff and topological dimensionalities. The experimental studies were conducted on corpora from national literatures in six languages – Russian, English, Kazakh, Kyrgyz, Tatar, and Uzbek – belonging to different typological groups. Vector representations of n-grams were formed using singular value decomposition of the context matrix, which allowed the dimensionality of embedding spaces to be varied without retraining the models.

The results revealed consistent differences in the intrinsic dimensionalities of semantic spaces of the studied languages and confirmed their multifractal nature. Interpretation of the findings suggests that the identified differences are due to the typological and structural features of the languages. The obtained estimates are robust to noise and changes in the dimensionality of the embedding space, ensuring the reproducibility of the results.

The practical significance of this work relates to the possibility of using intrinsic dimensionality as an engineering parameter in the design and optimization of natural language processing systems to reduce computational and resource costs

Keywords: intrinsic dimensionality, semantic spaces, graph algorithms, fractal structure, vector representations

Received 07.11.2025
Received in revised form 15.01.2026
Accepted date 22.01.2026
Published date 27.02.2026

***How to Cite:** Yerbolova, A. S., Kurmashev, I. G. (2026).*
Revealing intrinsic dimensionality patterns in semantic spaces of natural languages using graph algorithms.
Eastern-European Journal of Enterprise Technologies, 1 (2 (139)), 68–76.
https://doi.org/10.15587/1729-4061.2026.351509

1. Introduction

Modern natural language processing (NLP) methods rely heavily on vector embedding spaces to represent words, n-grams, as well as more complex linguistic objects. Such representations underlie most modern algorithms for semantic analysis, information retrieval, and machine translation. Despite their high practical effectiveness, the choice of semantic space dimensionality is in most cases heuristic and determined by architectural limitations of the models or established empirical standards. This leads to overparameterization of models, increased computational costs, and decreased scalability of NLP systems.

From a substantive perspective, the semantic space of language, which reflects the structure of meanings and their interrelations, is approximated in practical NLP systems using vector embedding spaces. However, the geometric properties of this space, in particular its intrinsic dimensionality, are typically not considered in the design and analysis of embedding models. Meanwhile, the results of experimental and theoretical studies show that the sets of vector representations of linguistic units have non-integer, fractal dimensionalities, making their description in terms of classical smooth manifolds incorrect.

From an engineering perspective, the problem of estimating the intrinsic dimensionality of a language's semantic space is closely related to the problem of optimally choosing the dimensionality of embedding representations. The use of vectors with dimensionalities of hundreds and thousands of components, typical of modern neural network models (such as ELMo [1], BERT [2], GPT-3 [3], and their derivatives), leads to significant computational and resource costs. Therefore, the search for quantitatively valid methods for determining the minimum sufficient dimensionality of an embedding space that preserves its structural and semantic properties is particularly relevant.

The specific nature of natural languages as complex self-organizing systems determines the heterogeneity of the geometric properties of their semantic spaces. Differences between languages an affect the intrinsic dimensionality of embedding representations, making its analysis in a cross-linguistic context a relevant scientific and applied task.

Thus, investigating the intrinsic dimensionality of semantic spaces in natural languages, as well as techniques for approximating them using embedding models, is a relevant scientific and applied task, important for the design, analysis, and optimization of modern natural language processing systems.

2. Literature review and problem statement

Existing research related to the analysis of language data can be roughly divided into two interrelated areas. The first area includes works that consider natural language as a holistic system and study its global structural and statistical properties. The second area focuses on works that study the intrinsic dimensionality of a given set of points.

The first area of research is based on the concept of language as a complex system governed by universal statistical laws. Thus, work [4] examines power laws of natural language, demonstrating that this class of distributions governs most fundamental linguistic processes. Continuing this line of thought, work [5] examines long-term correlations in natural languages, also considering language as a complex adaptive system and confirming the existence of stable statistical universals invariant to a specific language. Study [6] draws parallels between linguistic laws and biological systems, emphasizing the role of fundamental structural constraints that shape language. Despite the value of these results, it's worth acknowledging their key drawback: the focus is exclusively on the statistical properties of sequences of symbols and words, while the geometry of semantic spaces remains outside the scope of the analysis.

Separately, we note papers [7, 8], which examine natural language as a dynamic system governed by a self-organizing critical system. Specifically, in [7], literary texts are interpreted as an "avalanche" in semantic space. However, despite the conceptual significance of this approach, it does not allow for quantitative estimates of the geometric characteristics of semantic spaces, such as their intrinsic dimensionality.

In the second line of research on intrinsic dimensionality estimation algorithms, it is important to refer to [9], which proposed an axiomatic approach to the concept of intrinsic dimensionality of a dataset based on the geometry of multidimensional structures. That paper introduces the necessary conditions for considering a function as intrinsic dimensionality, drawing on the concept of distance between metric spaces proposed in [10]. These results provide a rigorous mathematical context; however, they are not adapted for working with empirical language data.

Existing methods for estimating intrinsic dimensionality based on representative point samples can be roughly divided into three main classes. The first class includes methods based on the analysis of strange attractor trajectories and time series containing the set of interest [11, 12]. The second class is based on graph methods and the analysis of proximity structures, such as minimum spanning trees [13]. The third class uses methods of topological data analysis, in particular persistent homology [14].

Despite the presence of modern methods for estimating intrinsic dimensionality and their consistent application to semantic spaces of language vector representations (embeddings), this approach cannot be considered systematic. Most studies focus either on the analysis of superficial character sequences or on methods for estimating intrinsic dimensionality that are not applicable to linguistic semantic spaces and are used primarily for abstract geometric objects. In particular, the question of whether there are consistent differences in the intrinsic dimensionality of semantic spaces of languages from different typological groups, as well as whether such differences can be detected using robust graph algorithms, remains open.

Thus, our review of the literature [4–14] shows that the problem of quantitatively assessing the intrinsic dimensionality of semantic spaces of natural languages, as well as its comparison in an interlingual context, remains insufficiently studied. This predetermines the need for targeted research focused on the application of graph methods for assessing intrinsic dimensionality to linguistic embedding spaces.

3. The aim and objectives of the study

The aim of our work is to quantitatively estimate the intrinsic dimensionality of semantic spaces of natural languages using graph algorithms and analyze cross-language differences. This will enable the use of intrinsic dimensionality as an engineering parameter for the informed selection and optimization of embedding representation dimensionality in natural language processing systems, thereby reducing computational and resource costs without losing the semantic structure of the data.

The formal problem statement includes the following steps:

– to construct vector models for each studied natural language (Kazakh, Russian, English, Tatar, Uzbek, and Kyrgyz) based on representative text corpora, generating sets of vector representations of all unique n-grams in embedding spaces of varying dimensionality;

– to estimate the intrinsic dimensionality of embedding spaces of fixed-length n-grams using graph dimensionality estimation algorithms;

– to perform a comparative analysis of the estimates of the Hausdorff and topological intrinsic dimensionality of the semantic embedding spaces of unigrams, bigrams, and trigrams of the languages under study and to test the hypothesis about the multifractal nature of these spaces.

4. Materials and methods

4. 1. The object and hypothesis of the study

The object of our study is semantic embedding spaces of natural languages, formed based on vector representations of n-grams of literary texts.

This study tests the hypothesis that the intrinsic dimensionality of linguistic semantic spaces is non-integer (fractal) and that its numerical values systematically vary between languages belonging to different typological groups. This assumption is tested under the following conditions:

1. Corpora of national literatures represent the basic semantic structure of language.

2. Vector representations obtained using singular value decomposition describe the geometry of semantic spaces.

3. Graph methods for estimating intrinsic dimensionality are applicable to high-dimensional linguistic semantic spaces obtained from linguistic data.

4. 2. Data and construction of vector representations (embeddings)

4. 2. 1. Description of text corpora

For our large-scale experiment, corpora of national literature in six natural languages (Russian, English, Kazakh, Kyrgyz, Tatar, and Uzbek) were used, as literary works are considered the stable core of natural language. All texts in the corpora were obtained from open sources, and each corpus consists of tens of thousands of literary works. For subsequent analysis, the total number of texts in the corpus, as well as the number of unique unigrams, bigrams, and trigrams, were recorded.

Before constructing the embeddings, each text underwent a preprocessing procedure, including tokenization, stopword removal, lemmatization, and normalization of proper names and numerals.

4. 3. Construction of vector representations of language units

A method based on the singular value decomposition (SVD) of the context matrix was used to construct vector representations of language units. SVD was implemented using the Golub-Kahan-Lanczos algorithm [15]. Rows of the matrix of left singular vectors were used as the final vector representations of words (embeddings). The main advantage of SVD was the ability to change the dimensionality of embeddings without retraining. By simply truncating the decomposition, we were able to obtain spaces of different dimensionalities from the same computational session, ensuring full comparability of experiments.

Vector representations for unigrams, bigrams, and trigrams were obtained by concatenating the embeddings of the words they comprised. Thus, for each language, three separate sets of vectors were built, corresponding to n-grams of different lengths.

4. 4. Graph methods for estimating intrinsic dimensionality

To estimate the intrinsic dimensionality of linguistic objects, topological (Lebesgue dimensionality) and Hausdorff dimensionality metrics are used. The result of the first metric is an integer characteristic, while the second metric allows for fractional values, which is essential for analyzing fractal structures [14, 16]. The primary focus is on the formation and analysis of sets of vector embeddings of n-grams obtained from text in natural languages such as Kazakh, Russian, English, Tatar, Uzbek, and Kyrgyz.

As a mathematical foundation, we rely on an axiomatic approach to defining intrinsic dimensionality as a functional on a space with a metric ρ and measure V. This approach requires three key properties that ensure the correctness of estimates [3, 13]:

1) concentration property: the growth of dimensionality in a family of spaces (X_d) should correlate with the fulfillment of the Levy property;

2) continuity (stability): the dimensionality estimate should change smoothly as the spaces converge in the Gromov metric $d_{conc}(X_d,X)\to 0$, i.e., be robust to small perturbations of the data: $\delta(X_d)\to\delta(X)$;

3) normalization property: for standard objects, such as the d-dimensional unit sphere S^d, the estimate should yield a value asymptotically close to d.

To estimate the intrinsic dimensionality of embedding sets, nonparametric graph algorithms that utilize the properties of the minimum spanning tree (MST) were chosen due to their applicability to high-dimensional data and the absence of a priori assumptions about the geometric shape of the data medium.

The first algorithm is used to estimate the Hausdorff (fractal) dimensionality of a point set and is based on an analysis of the scaling dependence of the total edge lengths of the minimum spanning tree as the sample size changes. For each point set (vector representation), a minimum spanning tree is constructed for subsamples of varying sizes, after which the corresponding graph statistics are calculated.

The second algorithm estimates the topological (integer) dimensionality through the degree distribution of the MST vertices. To establish the dependence of the graph statistics on dimensionality, a Monte Carlo calibration procedure is first performed on synthetic samples in spaces of known dimensionality.

4. 5. Validation procedures and experimental conditions

Formal conceptual analysis was used to independently verify the correctness of the relative ordering of the resulting intrinsic dimensionality estimates. This approach was used solely as a validation tool and was not considered the primary tool for quantitative assessment.

All computational experiments were performed under identical software and hardware conditions. Identical algorithms, parameters, and data processing procedures were used for all languages, n-gram orders, and embedding space dimensionalities, ensuring reproducibility and comparability of our results.

5. Results of the estimation of the internal dimensionality of semantic spaces of natural languages

5. 1. Formation of vector representations

Based on the corpora of national literature for each language under study, vector representations of unigrams, bigrams, and trigrams were constructed using the singular value decomposition method of the context matrix.

The original corpus of texts $I = (\Omega_1, \ldots, \Omega_L)$ and unique $\aleph = (\lambda_1, \ldots, \lambda_M)$ are used to construct a weighted matrix $W = (w_{ij})$ of dimensionality $M \times L$. Its elements are defined as

$$w_{i,j} = \left(1-\varepsilon_i\right)\frac{k_{i,j}}{\sum_{i'\in\Omega_j} k_{i',j}}, \tag{1}$$

where $k_{i,j}$ is the number of occurrences of word λ_i in text Ω_j, and ε_i is the normalized entropy of the distribution of the word in the corpus, calculated using the formula:

$$\varepsilon_i = -\frac{1}{\log L}\sum_{j=1}^{N}\frac{k_{i,j}}{\tau_i}\log\frac{k_{i,j}}{\tau_i}, \tag{2}$$

$$\tau_i = \sum_{j=1}^{N} k_{i,j}. \tag{3}$$

Thus, matrix W reflects words that are frequently encountered in a particular text but are not common to the entire context. Singular value decomposition is applied to matrix W [4, 17]

$$W \simeq W' = U\Lambda V^T, \tag{4}$$

where U and V^T are $M \times d$ and $d \times L$ matrices, respectively, and Λ is a diagonal matrix with singular values. The hyperparameter d specifies the amount of information stored for reconstruction W'. The SVD implementation is performed using the Golub-Kahan-Lanczos algorithm [15], and the rows of matrix U provide vector representations of words $\aleph$ of dimensionality d.

The advantage of this approach is the ability to compute embeddings for any smaller value $d_2 < d_1$ using the first d_2 components from the vectors computed for d_1 [12]. This allows for computational experiments to be conducted for multiple dimensions of embedding vectors without significant training costs, unlike other methods, such as deep neural networks. The vector representation of a n -gram is obtained by concatenating the embeddings of its constituent words in the order they appear.

5. 2. Estimating intrinsic dimensionality using graph algorithms

5. 2. 1. Graph methods for estimating intrinsic dimensionality (Schweinhart and Brito algorithms)

To estimate the intrinsic dimensionality of n-gram sets in the studied natural languages, two nonparametric algorithms based on constructing a minimum spanning tree (MST) of the data graph were used [5, 13]. Graph methods were chosen due to their robustness to noise [13] and their ability to identify both local and global structural patterns, which is critical for working with noisy text data from open sources.

Method (Schweinhart). This method allows one to estimate the fractal (Hausdorff) dimensionality d^* of a point set representing data in the embedding space. It is based on a theoretical result linking the sum of weighted edges of an MST with the dimensionality of the support of the measure from which the sample was obtained [18].

Let (x_1, ..., x_l) be a finite sample of points in $\mathbb{R}^d$, and T(x_1, ..., x_l) be its minimum spanning tree. For a given parameter $\alpha > 0$, the statistic is calculated

$$E_\alpha^0\left(x_1,\ldots,x_\ell\right)=\sum_{e\in T_\ell\left(\{x_\ell\}_{\ell\in\mathbb{N}}\right)}|e|^\alpha, \tag{5}$$

where $|e|$ is the Euclidean length of edge e in the tree $T(x_1,..., x_l)$.

Theoretically, for a sample of d^*, a regular Lefors measure, for $0 < \alpha < d^*$, then:

$$\text{const}_1\le\frac{E_\alpha^0\left(x_1,\ldots,x_\ell\right)}{\ell^{1-\alpha/d^*}}\le\text{const}_2, \tag{6}$$

$$\frac{\ln\left(E_\alpha^0\left(x_1,\ldots,x_\ell\right)\right)}{\ln\left(\ell\right)}\to\frac{d^*-\alpha}{d^*}, \tag{7}$$

$l \to \infty$; const_1 *u* const_2 are positive and independent of the constant l.

Based on this theorem, it becomes possible to estimate the fractal dimensionality $\hat{d}_{Schw}$ of a geometric structure represented by a point set $\left\{x_\ell\right\}_{\ell\in\mathbb{N}}, x_\ell\in\mathbb{R}^d$. The practical estimation procedure involves the following steps: for a fixed parameter α, the E_α^0 statistic is calculated for successively increasing subsamples of varying sizes l. This generates a set of value pairs that is used to solve the regression problem described by the following model

$$\ln\left(E_\alpha^0\left(x_1,\ldots,x_\ell\right)\right)=\ln\left(\psi\left(\alpha,d^*\right)\right)+\ln\left(\ell\right)\frac{d^*-\alpha}{d^*}. \tag{8}$$

To find the unknown parameter d^*, the nonlinear least squares (LSM) method is applied to these data. The resulting estimate $\hat{d}_{Schw}$ is interpreted as an approximate value of the upper bound of the Hausdorff dimensionality of the set under study [18], based on

$$\frac{d^*-\alpha}{d^*}\approx\frac{\hat{d}_{Schw}-\alpha}{\hat{d}_{Schw}}.$$

Brito's Method. The second method for estimating intrinsic dimensionality is based on the analysis of the structural properties of a minimum spanning tree (MST), namely, the distribution of its vertex degrees [13]. It has been found that graph statistics such as the mean squared vertex degree exhibit high robustness to noise and the growth of the embedding space dimensionality, making them suitable for estimating the topological (integer) dimensionality of the d_T estimator [19].

Computational procedure for estimation:

1. Calculation of key statistics. For a sample of points (x_1, ..., x_n) in space X_d, an MST is constructed. The dimensionality-sensitive statistic is the mean squared vertex degree of this tree

$$M_\ell\left(d\right):=\frac{1}{\ell}\sum_{x_i\in T\left(\{x_\ell\}_{\ell\in\mathbb{N}}\subset X_d\right)}\left(\deg\left(x_i\right)\right)^2. \tag{9}$$

The limiting distribution of a given quantity as $l \to \infty$ depends on the intrinsic dimensionality d^* [13].

2. Calibration using the Monte Carlo method. To calibrate the dependence of distribution M_l on dimensionality i, the uniform distribution in the unit hypercube is used. For each candidate i (in a range of reasonable dimensionalities), the Monte Carlo method generates K samples of size l. Using these synthetic data, the parameters of the normal prior distribution are estimated $f_{prior}\left(M_\ell|d=i\right)\sim N\left(\hat{\mu}_i,\hat{\sigma}_i^2\right)$:

$$\hat{\mu}_i:=\frac{1}{K}\sum_{j=0}^{K}M_\ell\left(i;\left\{x_\ell\right\}_j\sim U\left(0_i,1_i\right)\right), \tag{10}$$

$$\hat{\sigma}_i^2:=\frac{\ell}{K-1}\sum_{j=0}^{K}\left(M_\ell\left(i;\left\{x_\ell\right\}_j\sim U\left(0_i,1_i\right)\right)-\hat{\mu}_i\right)^2. \tag{11}$$

3. Bayesian inference. For a real sample of size ℓ', the observed value of the statistic M'_ℓ. is calculated. Using Bayes' theorem and the prior distributions obtained in the previous step, the posterior discrete probability distribution for the true dimensionality is calculated

$$p\left\{\hat{d}=i\right\}=p\left(i|M'_\ell\right)=\frac{N\left(M'_\ell;\hat{\mu}_i,\frac{\hat{\sigma}_i^2}{\ell'}\right)}{\sum_j N\left(M'_\ell;\hat{\mu}_j,\frac{\hat{\sigma}_j^2}{\ell'}\right)}. \tag{12}$$

4. Obtaining the final estimate. The estimate of the intrinsic (topological) dimensionality $\hat{d}_{BQY}$ is calculated as the mathematical expectation $E\left[\hat{d}\right]$ of this posterior distribution, rounded to the nearest integer:

5. Calculating the estimate of intrinsic dimensionality $\hat{d}_{BQY}$ as the mathematical expectation of $E\left[\hat{d}\right]$, rounded to the nearest integer

$$\hat{d}_{BQY}\left(M'_n\right)=\text{round}\left(\sum_i p\left(i|M'_\ell\right)*i\right). \tag{13}$$

It has been proven that $\hat{d}_{BQY}$ converges to d_T as K, l, $l'\to\infty$ [3].

A method based on formal conceptual analysis [20–22] was used as a tool for validating the order of the obtained dimensionalities. Although this method yields approximate estimates for real data, it allows for independent verification of the correctness of the order of magnitudes obtained using graph methods.

5. 2. 2. Numerical estimates of the intrinsic dimensionality of standard geometric objects

This study obtained numerical estimates of the intrinsic dimensionality for standard geometric objects with a predetermined dimensionality, including smooth manifolds, monofractal, and multifractal sets embedded in Euclidean spaces of varying dimensionalities. In [19], these algorithms were tested on manifolds, fractal, and multifractal sets embedded in spaces of varying dimensionalities, allowing the obtained results to be used as a benchmark for interpreting linguistic data.

The Schweinhart algorithm provides both point and interval estimates of the intrinsic (Hausdorff) dimensionality. To improve the reliability of the estimates, the dependence of the obtained dimensionality on parameter α was taken into account. Only those values for which the confidence interval of the estimate did not exceed the specified threshold γ were included in the analysis; other estimates were considered uninformative and were discarded. Only acceptable parameter values were used in the subsequent analysis.

Numerical experiments reported in [19] revealed characteristic patterns of dependence of the intrinsic dimensionality estimate on parameter α for various classes of objects. For smooth manifolds and monofractal sets, stable horizontal sections corresponding to integer and fractional values of the intrinsic dimensionality were observed. Multifractal structures are characterized by an arc-shaped dependence, reflecting the presence of a range of scales and the heterogeneity of the geometric structure.

In all the cases examined, a divergence in confidence intervals was noted with increasing α, which is associated with a violation of the applicability condition of Steele's theorem ($0 < \alpha < \partial^*$). Analysis of dependences $\hat{d}_{Schw}$ for standard multifractals revealed the following characteristic features: a non-integer value of the intrinsic dimensionality, as well as the presence of local maxima and minima, interpreted as upper and lower estimates of the true dimensionality. Linear dependences at large values of α were considered artifacts of the method and excluded from the interpretation, which is consistent with the results in previous studies [18].

The Brito algorithm was tested in a similar manner. The results were represented as scatterplots, where the sample size is plotted on the abscissa and the topological dimensionality estimate is plotted on the ordinate. The experiments confirmed that the accuracy of the estimates decreases with a high ratio of the intrinsic dimensionality of the object to the dimensionality of the embedding space. Moreover, an increase in the sample size led to stable convergence of estimates, which made it possible to use this algorithm as an additional tool for monitoring the results obtained by the Hausdorff method.

5. 3. Estimation of Hausdorff and topological intrinsic dimensionalities of semantic embedding spaces of natural languages

To estimate the intrinsic dimensionality of semantic spaces of natural languages represented as embedding spaces, vector representations of unigrams, bigrams, and trigrams were generated using national literature corpora. All texts were obtained from open sources and underwent a standard preprocessing procedure, including stopword removal, tokenization, and lemmatization.

For Russian, the corpus included 6,429 texts containing 103,952 unique unigrams, 14,775,439 bigrams, and 147,533,142 trigrams. For English, 11,052 texts were used, including 94,087 unigrams, 9,490,603 bigrams, and 39,173,019 trigrams. Corpora with similar structures were created for the Kazakh, Kyrgyz, Tatar, and Uzbek languages.

To assess the internal dimensionality of linguistic fractal structures, several values of the embedding space dimensionality were chosen: $d = \{5, 10, 15\}$. Research on the application of the entropy-complexity pair analysis method [12] served as an important reference point. According to the results of our study, for $n = 2$ and $n = 3$, acceptable values of dimensionality d do not exceed 15.

For the Schweinhart algorithm, the following computational procedure was performed for each language:

1) a sequence of spanning trees was generated, with n ranging from 1e + 5 to the size of the dataset;

2) for each n, with $\boldsymbol{\alpha}$ ranging from 1e-4 to 10 with a step of $s \approx 0.1$, the regression parameters were estimated to calculate $\hat{d}_{Schw}$;

3) inadmissible values were discarded according to the following criteria:

a) confidence intervals for the regression line are too large (greater than $\boldsymbol{\gamma}$);

b) confidence intervals for parameter $\frac{\hat{d}-\alpha}{\hat{d}}$ are too large (greater than $\boldsymbol{\gamma}$);

4) the minimum and maximum estimates $\hat{d}_{Schw}$ for all admissible $\boldsymbol{\alpha}$ were selected as the final estimates.

For the Brito algorithm, for a given language:

1) forall$d = i \in 2 \ldots 15$,samples $\left\{x_1^j, \ldots, x_{1e+6}^j\right\}_{j=1..100}, x \sim U\left(0_i, 1_i\right)$, were generated, where 0_i,1_i are the vectors of zero and one, respectively, in i-dimensional space;

2) the $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ distribution parameters were estimated for each i;

3) estimates $\hat{d}_{BQY}\left(\left\{x_1, \ldots, x_\ell\right\}\right)$. were calculated for each l.

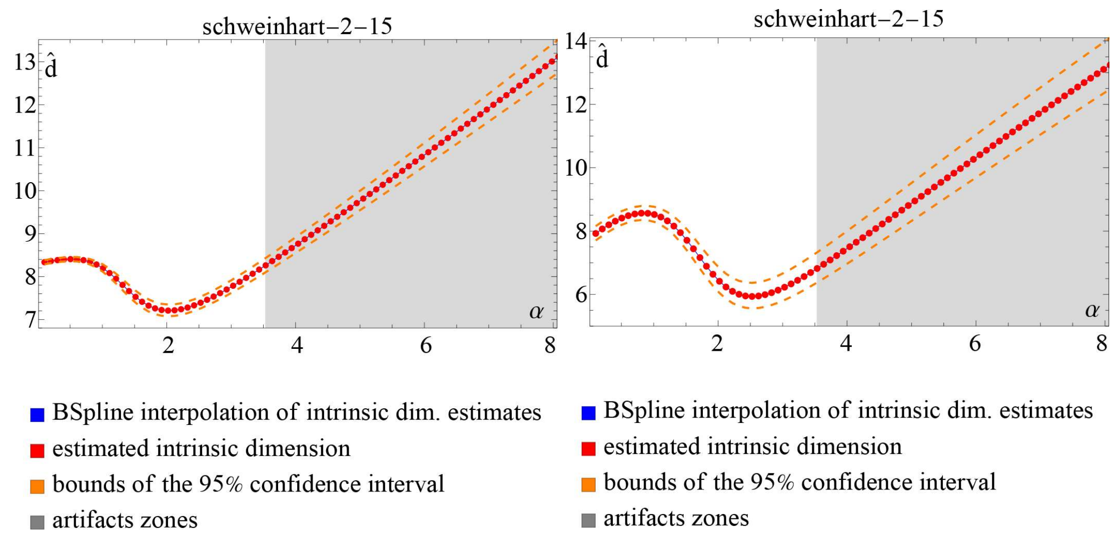
Typical dependences of intrinsic dimensionality estimates on parameter $\boldsymbol{\alpha}$, obtained using the Schweinhart algorithm for the languages under study, are shown in Fig. 1, 3, 5. Similar dependences of topological dimensionality estimates on sample size, obtained using the Brito algorithm, are shown in Fig. 2, 4, 6.

Numerical estimates of the internal dimensionality of the semantic spaces of the studied languages, obtained using SVD representations and graph methods, are given in Table 1, which contains the results for unigrams, bigrams, and trigrams for each language.

Table 1

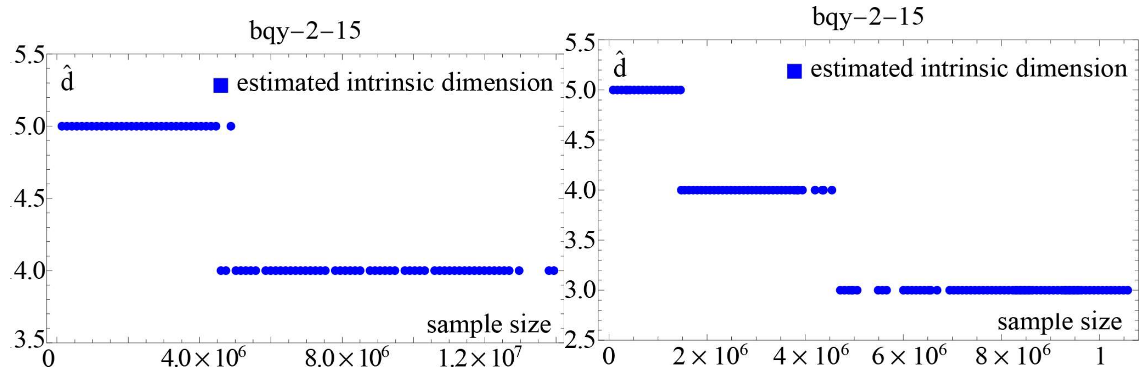
Estimating the intrinsic dimensionality of SVD representations of Indo-European languages

Language	n	d	$\hat{d}_{BQY}$	max $\hat{d}_{Schw}$	min $\hat{d}_{Schw}$	α
Kazakh	1	15	7–9	8.78	5.77	0.1–1.61
	2	15	8–10	6.52	6,08	0.1–5.05
	3	15	8–10	6.67	5.89	0.1–5.05
Kyrgyz	1	15	9–10	8.85	8.51	0.1–1.31
	2	15	9–11	6.51	5.19	0.1–4.04
	3	15	10–11	6.62	6,03	0.1–6.46
Tatar	1	15	8–10	8.19	5.76	0.1–1.61
	2	15	9–10	6.63	5.64	0.1–4.94
	3	15	9–10	6.72	6.41	0.1–4.24
Uzbek	1	15	7–10	7.95	6.45	0.1–1.71
	2	15	7–10	6.32	5.48	0.1–5.05
	3	15	8–10	6.56	6.38	0.1–4.24
Russian	1	15	11–12	10.62	9.64	0.1–1.01
	2	15	13–15	8.40	7.20	0.1–3.53
	3	15	14–15	8.59	7.96	0.1–4.34
English	1	15	10–11	10.39	8.81	0.1–1.01
	2	15	10–15	8.56	5.95	0.1–3.53
	3	15	13–15	9.29	5.97	0.1–1.91



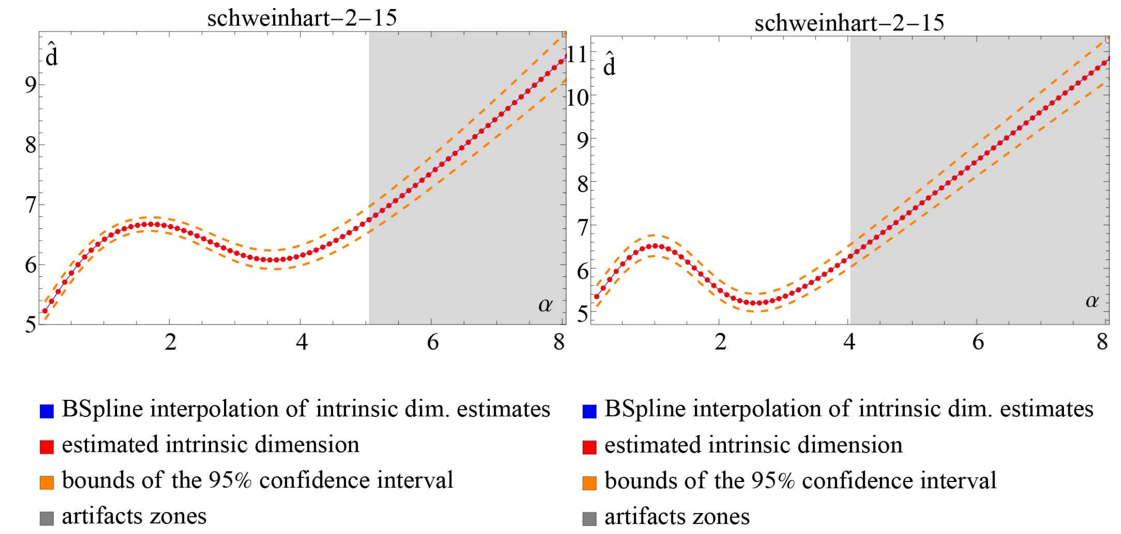
a *b*

Fig. 1. Dependence of the estimates obtained using the Schweinhart graph-based algorithm on the scaling parameter for vector representations of national literature: *a* — Russian language; *b* — English language



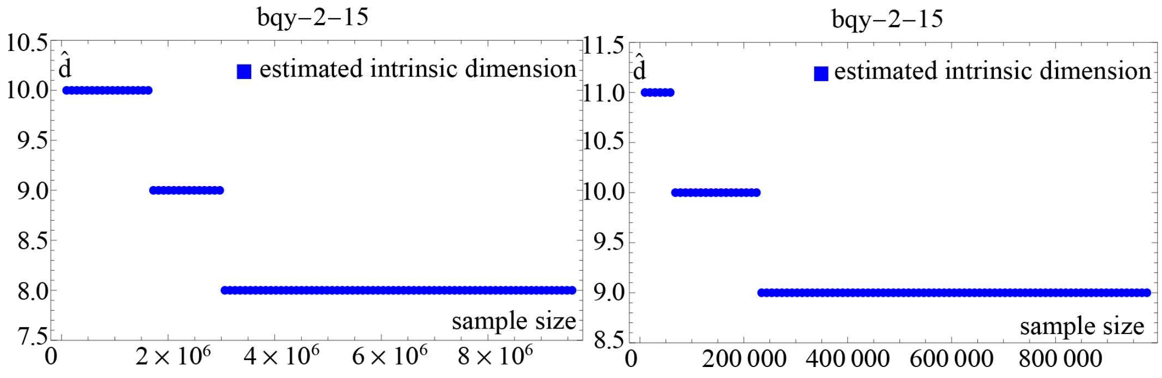
a *b*

Fig. 2. Dependence of the estimates obtained using the Brito graph-based algorithm on the sample size for vector representations of national literature: *a* — Russian language; *b* — English language



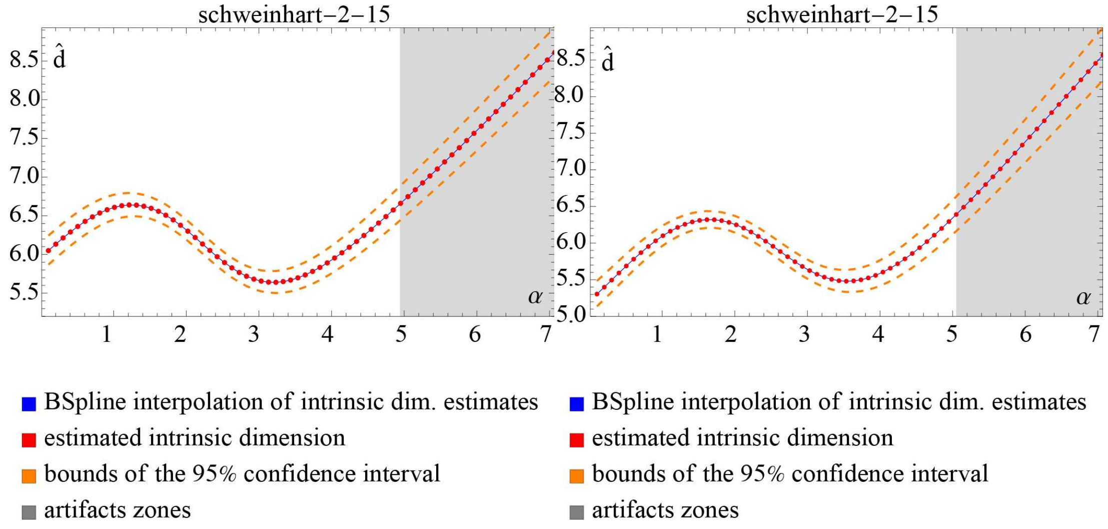
a *b*

Fig. 3. Dependence of the estimates obtained using the Schweinhart graph-based algorithm on the scaling parameter for vector representations of national literature: *a* — Kazakh language; *b* — Kyrgyz language



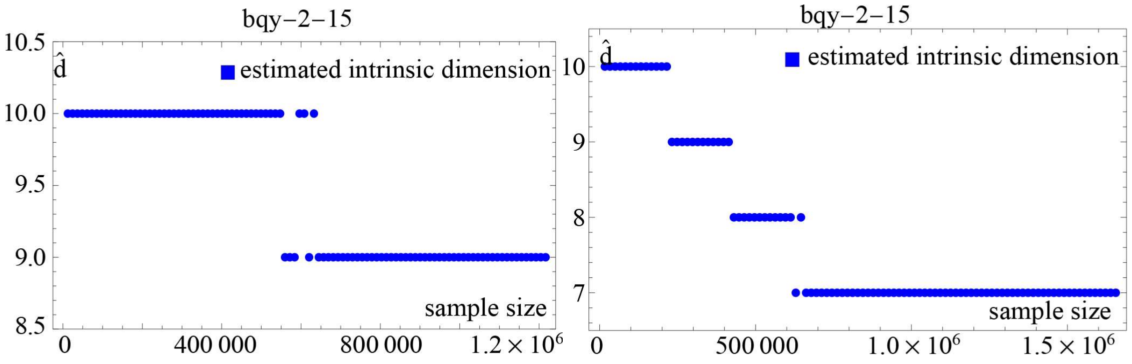
a *b*

Fig. 4. Dependence of the estimates obtained using the Brito graph-based algorithm on the sample size for vector representations of national literature: *a* – Kazakh language; *b* – Kyrgyz language



a *b*

Fig. 5. Dependence of the estimates obtained using the Schweinhart graph-based algorithm on the scaling parameter for vector representations of national literature: *a* — Tatar language; *b* — Uzbek language



a *b*

Fig. 6. Dependence of the estimates obtained using the Brito graph-based algorithm on the sample size for vector representations of national literature: *a* — Tatar language; *b* — Uzbek language

Analysis of the data given in Table 1 reveals that for all languages studied, the estimates of the intrinsic dimensionality of semantic spaces are significantly lower than the dimensionality of the embedding space. Non-integer values of the Hausdorff dimensionality indicate the fractal nature of the sets of vector representations of linguistic units.

6. Discussion of the results of estimating the intrinsic dimensionality of semantic embedding spaces of natural languages

The estimates of the intrinsic dimensionality of semantic embedding spaces in natural languages reported in our study demonstrate consistent and reproducible differences between the language groups studied. According to the results, the dimensionality of unigrams in Indo-European languages is 10.62 for Russian and 10.39 for English, while for bigrams and trigrams, the estimates decrease to values of 8.40–9.29. For the Altaic languages (Kazakh, Kyrgyz, Tatar, and Uzbek), the intrinsic dimensionality ranges from 7–9, while for bigrams and trigrams, it settles around values of 6–6.7.

The plots of the estimates as a function of algorithm parameters, shown in Fig. 1–6, for both methods (Schweinhart and Brito) have a pronounced multifractal shape, indicating a heterogeneous geometric structure of the semantic embedding spaces of natural languages. The lower values of the intrinsic dimensionality of the Altaic languages indicate a more compact internal organization compared to Indo-European languages, which have greater structural complexity.

The main feature of the proposed approach is the application of nonparametric graph methods for estimating the intrinsic dimensionality of linguistic semantic spaces. Unlike other approaches, such as neural network models, this approach allows for varying the dimensionality of the embedding space without retraining the model while maintaining the stability of the estimates. Our results are consistent with the estimates of the intrinsic dimensionality for fractal and multifractal sets presented in [18, 19], confirming the validity of the method.

This paper has two major limitations. First, the intrinsic dimensionality estimates are sensitive to corpus characteristics such as size and genre. Second, the calculation results depend on the algorithm parameter settings.

Further research plans to expand the range of languages included in various typological groups and to use other vector representation algorithms.

7. Conclusions

1. As a result of our large-scale study, vector representations (embeddings) were developed for unigrams, bigrams, and trigrams in six different languages using singular value decomposition (SVD). The resulting embedding spaces allow analysis for various dimensionalities without the need for retraining the models, reducing computational costs and ensuring consistency of results across different languages and n-gram types. It was shown that the underlying geometric structure of the semantic spaces remains stable under this operation, which can be explained by the linear nature of SVD and its resistance to truncation.

2. Using graph algorithms, robust estimates of both the Hausdorff and topological intrinsic dimensionalities of the semantic spaces were obtained. For Indo-European languages, the obtained estimates of the Hausdorff dimensionality are in the range of 10.4–10.6 for unigrams, and for bigrams and trigrams, the dimensionality estimates decrease to 8.4–9.3. For the Altaic languages, these values are significantly lower: 7–9 for unigrams and approximately 6–6.7 for bigrams and trigrams. These graph algorithms are highly robust to noise and parameter variations, distinguishing the proposed approach from conventional methods. The observed differences between language families are due to varying degrees of structural complexity and the manifestation of multifractal properties in their semantic spaces.

3. A cross-linguistic comparative analysis revealed statistically significant differences in intrinsic dimensionality between language families. It was found that the semantic spaces of Indo-European languages are characterized by higher intrinsic dimensionality than those of the Altaic languages. Our results are consistent with existing concepts of language as a complex system and also confirm the multifractal nature of linguistic semantic spaces. These findings substantiate the use of intrinsic dimensionality as one of the engineering parameters when choosing the dimensionality of embedding representations. In particular, it is shown that with the dimensionality of the embedding space $d = 15$, the intrinsic Hausdorff dimensionality of semantic spaces is about 6–7 for Altaic languages and 8–10 for Indo-European languages, which indicates the possibility of reducing the dimensionality of embedding vectors by 1.5–2 times without losing their structural and semantic properties.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available, either in numerical or graphical form, in the main text of the manuscript.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technologies when creating the current work.

Authors' contributions

Assel S. Yerbolova: Conceptualisation, Methodology, Writing – Original Draft, Formal analysis, Investigation, Visualization, Writing – Review and Editing; **Ildar G. Kurmashev:** Conceptualisation, Methodology.

References

1. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2227–2237. https://doi.org/10.18653/v1/n18-1202

2. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P. et al. (2020). Language Models are Few-Shot Learners. arXiv. https://arxiv.org/abs/2005.14165
4. Dębowski, Ł. (2020). Information Theory Meets Power Laws. John Wiley & Sons. https://doi.org/10.1002/9781119625384
5. Tanaka-Ishii, K. (2021). Language as a Complex System. Statistical Universals of Language, 19–30. https://doi.org/10.1007/978-3-030-59377-3_3
6. Semple, S., Ferrer-i-Cancho, R., Gustison, M. L. (2022). Linguistic laws in biology. Trends in Ecology & Evolution, 37 (1), 53–66. https://doi.org/10.1016/j.tree.2021.08.012
7. Gromov, V. A., Migrina, A. M. (2017). A Language as a Self-Organized Critical System. Complexity, 2017, 1–7. https://doi.org/10.1155/2017/9212538
8. Malinetsky, G. G., Potapov, A. B. (2000). Sovremennye problemy nelineinoi dinamiki. Moscow: Editorial URSS.
9. Pestov, V. (2007). Intrinsic dimension of a dataset: what properties does one expect? 2007 International Joint Conference on Neural Networks, 2959–2964. https://doi.org/10.1109/ijcnn.2007.4371431
10. Gromov, M. (2007). Metric Structures for Riemannian and Non-Riemannian Spaces. Birkhäuser, 586. https://doi.org/10.1007/978-0-8176-4583-0
11. Kantz, H., Schreiber, T. (2003). Nonlinear Time Series Analysis. https://doi.org/10.1017/cbo9780511755798
12. Panda, S. K., Nagy, A. M., Vijayakumar, V., Hazarika, B. (2023). Stability analysis for complex-valued neural networks with fractional order. Chaos, Solitons & Fractals, 175, 114045. https://doi.org/10.1016/j.chaos.2023.114045
13. Brito, M. R., Quiroz, A. J., Yukich, J. E. (2013). Intrinsic dimension identification via graph-theoretic methods. Journal of Multivariate Analysis, 116, 263–277. https://doi.org/10.1016/j.jmva.2012.12.007
14. Adams, H., Aminian, M., Farnell, E., Kirby, M., Mirth, J., Neville, R. et al. (2020). A Fractal Dimension for Measures via Persistent Homology. Topological Data Analysis, 1–31. https://doi.org/10.1007/978-3-030-43408-3_1
15. Golub, G., Kahan, W. (1965). Calculating the Singular Values and Pseudo-Inverse of a Matrix. Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis, 2 (2), 205–224. https://doi.org/10.1137/0702016
16. Bellegarda, J. R. (2007). Latent Semantic Mapping. Latent Semantic Mapping: Principles & Applications, 9–13. https://doi.org/10.1007/978-3-031-02556-3_2
17. Kalman, D. (1996). A Singularly Valuable Decomposition: The SVD of a Matrix. The College Mathematics Journal, 27 (1), 2–23. https://doi.org/10.1080/07468342.1996.11973744
18. Schweinhart, B. (2020). Fractal dimension and the persistent homology of random geometric complexes. Advances in Mathematics, 372, 107291. https://doi.org/10.1016/j.aim.2020.107291
19. Steele, J. M. (1988). Growth Rates of Euclidean Minimal Spanning Trees with Power Weighted Edges. The Annals of Probability, 16 (4). https://doi.org/10.1214/aop/1176991596
20. Gromov, V. A., Borodin, N. S., Yerbolova, A. S. (2024). A Language and Its Dimensions: Intrinsic Dimensions of Language Fractal Structures. Complexity, 2024 (1). https://doi.org/10.1155/2024/8863360
21. Kuznetsov, S. O., Gromov, V. A., Borodin, N. S., Divavin, A. M. (2023). Formal Concept Analysis for Evaluating Intrinsic Dimension of a Natural Language. Pattern Recognition and Machine Intelligence, 331–339. https://doi.org/10.1007/978-3-031-45170-6_34
22. Kuznetsov, S. O. (2009). Pattern Structures for Analyzing Complex Data. Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, 33–44. https://doi.org/10.1007/978-3-642-10646-0_4