Semantic and Sentiment Trajectories of Literary Masterpieces

Preprint · May 2023

DOI: 10.2139/ssrn.4457882

2 authors:

Vasilii A. Gromov
National Research University Higher School of Economics
62 PUBLICATIONS **252** CITATIONS
SEE PROFILE

Quynh Nhu Dang
National Research University Higher School of Economics
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Semantic and sentiment trajectories of literary masterpieces

Vasilii A. Gromov and Quynh Nhu Dang
HSE University

Abstract

The paper deals with semantic and sentiment trajectories of literary masterpieces (we used corpora of 12 languages of various language families), composed of individual embeddings or n-grams. We ascertain that, for all languages, semantic and sentiment trajectories are markedly chaotic: positive largest Lyapunov exponents; 'entropy-complexity' pairs belonging to the 'chaotic' area of the respective plane; the distinctive 'chaotic' drop of the number of false nearest neighbours at a particular value of an embedding dimension. These characteristics are utilised in order to develop a method to compare styles of an original masterpiece and its translations (to automatically assess translation quality).

Keywords. Chaotic time series; sentiment trajectories; semantic trajectories; natural language as a whole; quality of machine translation; literature masterpieces

Introduction

Recent advances in natural language processing (*NLP*) make it possible to map words of a natural language into a real number (sentiment analysis [1]) or a vector of real numbers (word-to-vec and other similar methods [2,3]). In turn, this fact allows us to consider a text of a literary masterpiece as a trajectory (a path) of a dynamical system in a one- or multi-dimensional phase space. We propose to use the term a sentiment trajectory (path) if we calculate its elements using sentiment analysis techniques, and the term semantic trajectory, if its elements are word embeddings. We feel that such an object of study is of fundamental interest in itself; moreover, one can utilise it as a tool to verify hypotheses of literature studies.

The present paper explores both semantic and sentiment trajectories of literary pieces for all languages considered. We ascertained that both semantic and sentiment trajectories are markedly chaotic, and calculated characteristics of the respective strange attractors. It seems to us that such characteristics of a given text, compared with the respective mean values averaged over the language, are distinctive of an author's style ('The style is the man himself' [de Buffon], if anything). This implies, among other things, that we may compare the characteristics of an original masterpiece and its translation in order to assess the quality of the translation automatically. It is worth stressing that recent advances in NLP permit us to solve the problem of machine translation in an efficient and effective way (see [2,3] and references therein). However, the problem of automatic quality assessment still requires a human as an expert, and the expert should be fluent in both languages and have a good taste in literature, which may appear to be a hard task for the two

arbitrarily chosen languages. We feel that if one compares the characteristics of the original and translated texts, one can relate the style of the original piece of literature and that of its translation. A Japanese writer A. Ryūnosuke, in his novel "Mensura Zoili", describes a fictitious country Zoili, whose citizens designed a peculiar device (Mensura Zoili, for that matter) able to estimate the importance of literary work, and its author's talent, objectively and accurately. The comparison of styles seems to make it possible to redesign Mensura Zoili for interpreters ☺.

The choice of languages is determined, apart from the availability of voluminous corpora of texts, by their qualitative difference in grammar structure and language families. We consider languages of the Indo-European languages (Russian, English, German, Norwegian, French, Romanian, Hindi), the North Caucasian languages (Tabasaran), the Uralic languages (Finnish), and the Austroasiatic languages (Vietnamese); a language isolate (Basque); an artificial language (Esperanto). The languages differ in the prevalence of inflexions (for example, Russian is an inflected language, whereas in English, inflexions are rather rare), in word order (Russian is characterised by flexible word order in a sentence, whereas word order in the English language is strict, and rare exceptions are constrained by stringent rules [4,5]), and other characteristics. S1 Table of Appendix A lists all languages considered and their characteristics.

It is highly reasonable to investigate not only a series of word embeddings, but also n-grams. If we consider several consecutive words, we can make a more profound judgement about the strange attractor behind the text and the series itself.

The paper proposes a novel method to analyse literary masterpieces. The method implies that one:

1. Transforms words of a natural language into either real numbers (words sentiment characteristics) or real vectors (position of the words in a semantic space);
2. Considers a literary masterpiece as a one- or multi-dimensional time series that describes dynamics of the text in a sentiment or semantic space;
3. Ascertains whether the series are chaotic;
4. Examines distributions of these characteristics over texts of the natural language in question;
5. Compares the characteristics of an original masterpiece and its translations to design a new measure to assess the quality of the translation for both human- and machine translated texts.

In the framework of the method, we attempt to answer the following research questions:

1. Are sentiment and semantic trajectories of literature masterpieces chaotic?
2. What is the relation between the characteristics of trajectories for a masterpiece and its translations?

The remainder of the paper is as follows. In the next section, we discuss related works; in the third one, we consider characteristics of sentiment and semantic trajectories, and methods to estimate them. The fourth section presents the results of

a large-scale simulation for various languages. The fifth section compares original masterpieces and their translations. The last section presents conclusions and future directions.

Related works

The approaches to investigate literary masterpieces by means of formal, mathematical methods, depend upon the 'size' of the object at hand, the coarse-graininess of the study. The lowest level is occupied by a single word and methods to embed it into a semantic space (see reviews [2,3,6] and references therein). They share the level with works on local relations among words, closely resembling those that give rise to the syntax of natural languages [7].

For the next level, the object of study is a large semantic object within the text of a masterpiece. Complex networks [8,9] that reflect relations among various large semantic objects are a good example of the approaches associated with this level. Elson et al. [10] examine networks constructed for British novels of the XIX century; Mac Carron and Kenna [11], and Kydros et al. [12], for myth and sagas; Waumans et al. [13], for novels by J. K. Rowling. Stella and Brede [14] employ several complex networks to study the interplay between phonological and semantic features (the ‘multiplexity’) of the English language.

The next level deals with the whole text of a masterpiece, a masterpiece in its entirety. Most papers here explore sentiment trajectories [1,15,16] (‘happiness time series’, ‘emotional time series’). The site hedonometer.org/books.html is a large source of such trajectories and allows for downloading a great number of them. Generally, this approach attempts to somehow estimate emotional characteristics the authors associate with personages and their groups. Reagan et al. [1] show that sentiment trajectories, on a large scale, follow a small number of possible patterns. (The authors rely on a corpus of British novels for this end.) Similar results may be found in [15,17]. Dodds et al. apply sentiment analysis [15] to texts of several natural languages to test (and corroborate, indeed) the Pollyanna hypothesis [18]; the hypothesis asserts that real-world languages are biased toward emotionally positive words. Min and Park [19] propose to explore a literary masterpiece by calculating the mean Sentiment Polarity Index for each chapter to ascertain a chapter ‘topic’; the authors verify their approach on the English translation of Victor Hugo's “Les Miserables”.

We feel that the text of a literary masterpiece is a sequence of meanings (senses), rather than a sequence of emotional impulses. Consequently, we believe that semantic trajectories are more important, in a sense, than sentiment ones. Unfortunately, to the best of the authors' knowledge, semantic trajectories are poorly explored in the literature. We may only point to the books that discuss not semantic trajectories, but rather trajectories of words or symbols [20,21]. Prof. K. Tanaka-Ishii, in his monograph ([21] and references therein), studies long-range correlations in the English and Japanese languages. (The latter seems to be able to provide the results closest to our subject matter due to the hieroglyphic nature of the Japanese

language – there is a kind of one-to-one correspondence between meanings and hieroglyphs). The author indicates that long-range correlations in the languages amount to 10-15 words, which is much larger than conventional n-grams.

To complete the picture, let us mention the paper [22,23]. The authors study a natural language as an integral whole and ascertain that it is a self-organised critical system, whereas a separate literature text is 'an avalanche' in a semantic space. The latter fact further reinforces the argument for considering a trajectory in a semantic space as a unified object.

Of fundamental importance is the application of the semantic trajectories studies as tools to automatically assess translation quality. To the best of our knowledge, automatically assessing a given translated text is still a scientific problem. Most methods are still based either on human judgements [24-27] or on the calculation of similarity measures (WER, PER, BLEU, BP, CDER, and others) between machine and reference translations, which, once again, still requires a human interpreter [28-30].

Most papers that assess a translated text without human intervention deal with a separate word or sentence, not with a text as a whole [31,32]. Luong et al. [32] combine features of various types (system-based, lexical, syntactic and semantic) to solve the problems of Word-level Confidence Estimation, that is, to assess the translation of a single word. The new generation of machine translation technology [33-36] makes it possible to deal with a sentence as the smallest unit.

Among available measures used to assess translation quality of the whole text automatically, let us highlight the measures of lexical diversity: volume (i.e., text length), rarity (i.e., frequency of words in the language), and variability (i.e., type-token ratio corrected for text length) [37]. Graesser et al. [38] introduced Coh-Metrix 3.0 to measure volume, text length, and sentence length. Rarity measures were calculated for the proportion of K3 and K4 words (i.e., third and fourth 1000 English word families of English). To assess variability, McCarthy and Jarvis [39] proposed to employ textual lexical diversity (MTLD). It is worthy to note that Wang et al., in their review of recent advances of machine translation [40] in the section *Challenges and future directions*, indicate 'new evaluation metrics are needed to evaluate what really matters as the first problem.

To sum up, the literature makes it possible to conclude that semantic trajectories have not come under the scrutiny of science yet. Second, automatic translation quality deals mainly with separate words and sentences, frequently relying on human judgment.

Methods to construct and explore semantic trajectories

Preprocessing of natural language texts

Preprocessing implies that, first, all non-literal symbols and capital letters are replaced, and second, words are lemmatised. Third, we identify all named entities: names and family names, organisations, place names, and so forth. The identified

entities are replaced by names of the appropriate categories; this allows for obtaining more informative embeddings. S1 Table of Appendix A describes the libraries used for the preprocessing for various languages.

A sentiment trajectory of a text

In the framework of sentiment analysis, each word is mapped into a float number, ranging from -1 to 1. The value -1 corresponds to words that evoke the most negative emotions (like 'evil', 'death', 'torment'); the value 1, to words that evoke the most positive emotions (like 'good', 'life', 'love'). Thereby, a literature masterpiece (or also any other text) is mapped into a one-dimensional time series, a sentiment trajectory [1] to be studied. We succeeded in finding tools to obtain sentiment trajectories only for Russian texts. In order to obtain sentiment trajectories for the Russian language, we use a Python library *dostoyevksy*. Unfortunately, for other languages, we failed to find a tool that made it possible to quickly construct sentiment trajectories in such a way that each word possessed non-zero sentiment characteristic.

A semantic trajectory of a text

In order to determine the semantic trajectory of a text, we employ two concepts: context and vocabulary. A context is a set of texts $\mathfrak{T} = (\Omega_1, \cdots, \Omega_N)$; a vocabulary, a set of words $\aleph = (\lambda_1, \cdots, \lambda_M)$ of the respective natural language. It implies that any trajectory of a masterpiece depends on the context and vocabulary used to shape a semantic space. For a given context $\mathfrak{T}$ and vocabulary $\aleph$ elements of a TF-IDF matrix $W = (w_{ij})$ is defined as $TF(i,j) \cdot IDF(i,\Lambda)$, where

$$TF(i,j) = \frac{n_{i,j}}{\sum_{i' \in \Omega_j} n_{i',j}},\ IDF(i,\Lambda) = \frac{M}{|\{j \in \Lambda : i \in \Omega_j\}|} \quad (1)$$

$n_{i,j}$ is the number of times the word i enters the document j. Thus, the largest values correspond to words frequently occurring in a particular document, but rarely occurring in all other documents of the context. Most elements of a TF-IDF matrix are zero since a word usually enters a few documents and is absent in all other texts.

The singular value decomposition (SVD) of a matrix W [41] is a product:

$$W \simeq W' = U\Lambda V^T \quad (2)$$

U and V^T in (2) are rectangular $M \times d$ and $d \times N$, matrices, respectively; Λ is a square $d \times d$ matrix. Here d is a hyperparameter that determines a share of information available in the original matrix to retain. Matrix W' is proved to approximate W in the best possible way in the sense of the L_2-norm. To calculate SVD for a given matrix, we employ the Golub-Kahan-Lanczos algorithm [42], implemented in the Python library *scipy*. In doing so, we obtain a matrix U that consists of a required number of orthogonal column vectors; the columns are embeddings (of a predefined size), which represent the respective words. Furthermore, if one calculates such embeddings for a given $d = d_1$, then one is able

to obtain embeddings for any $d = d_2 < d_1$, just truncating the original $d = d_1$ embeddings to the first d_2 components [43]. This fact allows us to drastically reduce the required computational resources, which is of pivotal importance for the problem at hand (we should carry out simulation for a broader range of d). It is noteworthy too that the overwhelming majority of other methods to compute the embeddings (for example, those based upon neural networks) suggests that embeddings are calculated anew for each d.

So, singular value decomposition makes it possible to represent [41] each word $\lambda_i \in \aleph$ as a real vector of size d. In turn, this makes it possible to represent a text Ω_j as d-dimensional time series. We regard this time series as a trajectory of a dynamical system in a d-dimensional semantic space, a semantic trajectory of a text Ω_j. Different semantic trajectories correspond to different d's.

How to ascertain if a series is chaotic?

Chaotic time series are a visiting card of complex systems; one can observe them in a broad range of natural and social phenomena: quantum biology [44], breathing control [45], solar activity [46,47], electricity production and consumption [48], Twitter topics popularity [49], etc. To test the hypothesis that a given time series, a sentiment or semantic trajectory, is chaotic, we locate its position on the 'entropy - complexity' plane [50] and calculate the largest Lyapunov exponent [51,52].

Entropy - MPR complexity plane

Martin, Plastino, and Rosso (*MPR*) [50] propose an efficient and effective approach to distinguish a chaotic time series, on the one hand, from a series generated by a simple deterministic system, and, from a purely stochastic series on the other (white noise, conventional or fraction Brownian motion etc.). The approach implies that, for a given time series $\{x_i\}$, we calculate two characteristics – entropy and complexity – and compare the position of this point with the lower and upper theoretical boundaries to ascertain time-series type. Martin, Plastino, and Rosso [50] design algorithms for a one-dimensional time series. The algorithm itself is readily extendible to the case of a multi-dimensional time series, but the algorithm used to construct the lower and upper boundaries should be recalculated.

The algorithm relies upon a probability distribution of ordinal patterns [53] constructed for a given time series. An ordinal pattern means that, for a given m-dimensional real vector, one constructs an $d-1$-dimensional binary vector such that its component is equal to 1, if the i-th component of the real vector is less or equal to its $i+1$-th component, and equal to 0 otherwise. To generalise this concept for elements of multi-dimensional time series, we compare the first components of these elements, then the second, and so on. Thus, the defined pairwise comparison yields $\left((n-1)!\right)^d$ possible types of the binary vectors. Hence, the algorithm to construct a

probability distribution of ordinal patterns for a given time series involves the following steps:

1. One defines n and makes a sample of $N - n + 1$ vectors $s_i = \left(x_{i-(n-1)}, x_{i-(n-2)}, \ldots, x_{i-1}, x_i\right)$, where x_i are elements of the series under study (each of them consists of m components).
2. For each s_i, one ascertains the ordinal pattern type that it belongs to.
3. For each ordinal pattern type, one estimates its probability P_i as a ratio of the number of times it emerges in the series to the size of the sample.

For this distribution, one calculates two characteristics: entropy and complexity. The former is the conventional Shannon entropy, normalised to its maximum value:

$$S[P] = -\sum_{i=1}^{n!} P_i \cdot ln\ (P_i) \quad (3)$$

$$H[P] = \frac{S[P]}{S_{max}}$$

$S_{max} = ln\ (n!) = S[P_e]$, where P_e is the uniform distribution. (Entropy achieves its maximum at this very distribution.)

The second characteristic, MPR-complexity is defined as:

$$C[P] = Q[P, P_e] \cdot H[P]. \quad (4)$$

Here, Q is the Jensen-Shannon divergence between the given and the uniform distributions:

$$Q[P, P_e] = Q_0 \cdot \left(S\left[\frac{P+P_e}{2}\right] - S\left[\frac{P}{2}\right] - S\left[\frac{P_e}{2}\right]\right), \quad (5)$$

where Q_0 is a normalisation constant.

The procedure outlined above maps a time series into a point of the entropy-complexity plane. The position of the point in relation to the lower and upper theoretical boundaries points to the type of the series in question. Namely, simple deterministic processes occupy the bottom left corner of the plane, stochastic processes, the bottom right corner, whereas chaotic (complex deterministic) processes occupy areas adjacent to the vertex of the upper curve [50]. Figure 1 exhibits the lower and upper theoretical boundaries (red solid curves) and chaotic (discs), simple deterministic processes (squares), and simple stochastic processes (triangles). The figure also exhibits points associated with typical benchmarks of all three groups (we have repeated the simulation presented in Ref. [50]): triangles correspond to stochastic processes (fractional Brownian motion); squares, to simple deterministic processes (sine function); and disks, to chaotic processes (the Lorenz series, the logistic map, the skew tent map, the Henon map, and others). In what follows, we shall normalise the characteristics x to the respective average values over the respective language:

$$x' = \frac{x - \mu}{\sigma} \quad (6)$$

Consequently, all points below lay inside the unit square.

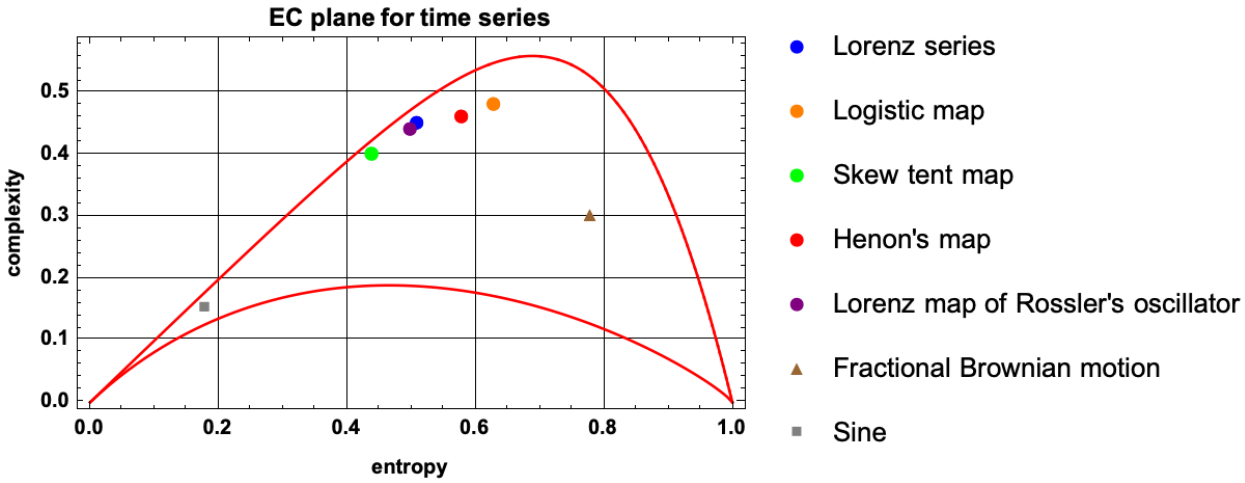


Fig 1. Stochastic (triangles), chaotic (discs), and simple deterministic (squares) on the entropy–complexity plane

The algorithm has many points in its favour: it is easily implementable, stable, and invariant with respect to non-linear transformation [50]. However, its performance depends strongly on values of n (the number of words), d (the size of word embedding), and N (the length of a series). As d and n increase, the number of ordinal pattern types grows exponentially. Consequently, for a relatively small time series (and all literary masterpieces produce such series), the majority of probabilities P_i are close to zero, and the distribution is close to the uniform. Therefore, the parameters should satisfy the following constraints: $N \gg n!$, $N \gg d!$. We think that the entropy-complexity pair reflects the true dynamics of a text if it belongs to the area of chaos, and we should account for only those values of n and d for which this is valid. It is hardly possible that the semantic trajectory of 'War and peace' (by Leo Tolstoy) reflects tossing a coin in a semantic space – evidently, we must attribute this to too small values of the embedding dimension and size of n-gram used. On the other hand, it is also hardly possible that the semantic trajectory of 'War and peace' is just mechanical repetition of the same ideas); thus, the shift to the area of simple deterministic process (we also cannot perceive that the semantic trajectory of 'War and peace' is attributable to too large values of these quantities.

The largest Lyapunov exponent

For chaotic systems, initially close trajectories tend to diverge exponentially with time. The average speed of the divergence is called the largest Lyapunov exponent (*LLE*). Its positive value reveals a chaotic system; negative, regular one. To estimate the LLE, we employ the algorithm proposed by Rosenstein [54]. The method is based upon the Poincaré recurrence theorem [55]; its corollary reads that for any time series that is long enough and for any section of this time series, we can find several sections similar to the given one, but not the same (its nearest neighbours). The average speed of divergence of time series sections and their

nearest neighbours (if any) gives the estimated LLE. For one-dimensional time series, we use the TISEAN package [56].

False nearest neighbours

For both the tests for chaos, discussed above, of fundamental importance is estimating the embedding dimension (the number of components of the z-vector) properly. To this end, it should satisfy Takens's theorem [57]. Its first condition requires that $p > 2d + 1$, that is, the size of a z-vector p should be larger than two by the dimension of the respective strange attractor plus one (Whitney theorem). Takens's theorem guarantees one-to-one correspondence between genuine dynamics, in the neighbourhood of a strange attractor, and observed dynamics, in a space of z-vectors. Since the dimension of a strange attractor d is usually unknown, scientists conventionally employ the false nearest neighbours (*FNN*) method to estimate it. It suggests calculating for every possible p, the number of nearest neighbours that are not nearest neighbours in the space of dimension $p + 1$. This implies that we append to each z-vector the observation that follows this z-vector in the series in question, and compare such extended z-vectors. For $p < 2d + 1$ the number of *FNN*s is small, whereas, when p achieves the threshold value, the number of *FNN*s drops rapidly, in a characteristic manner. This value of p gives a rough estimate of the dimension of the strange attractor. The method is discussed in great detail in Refs. [51,52].

Trajectories of literature masterpieces

In order to test the basic hypothesis that semantic trajectories of literature masterpieces are chaotic, we performed a large-scale simulation, based on corpora of English and Russian literature. S1 Table of Appendix A presents sizes of samples for all languages considered. We mostly employ prose masterpieces, since we need a rather large series to robustly estimate its characteristics. If we fail to get an access to the corpus of the national literature, we use the corpus of Wikipedia texts.

Sentiment trajectories

For the reasons indicated above, for sentiment trajectories, we present results for the Russian language only. Figure 2 exhibits a typical section of sentiment trajectories for the Russian language (a paragraph of "Crime and Punishment" by F. Dostoyevsky); Figure 3 demonstrates a distribution of the largest Lyapunov exponents for it. Evidently, the LLEs are strictly positive. This means that all sentiment trajectories are chaotic. The minimum LLE is 2.8e-5 (it is achieved on K. Fofanov "In grief"); the maximum LLE is 9.8e-2 (it is achieved on A. Pushkin's "Songs of the Western Slavs"); the average LLE is 3.17e-2 (it is achieved on S. Chorny "Politicon and Emigrant County").

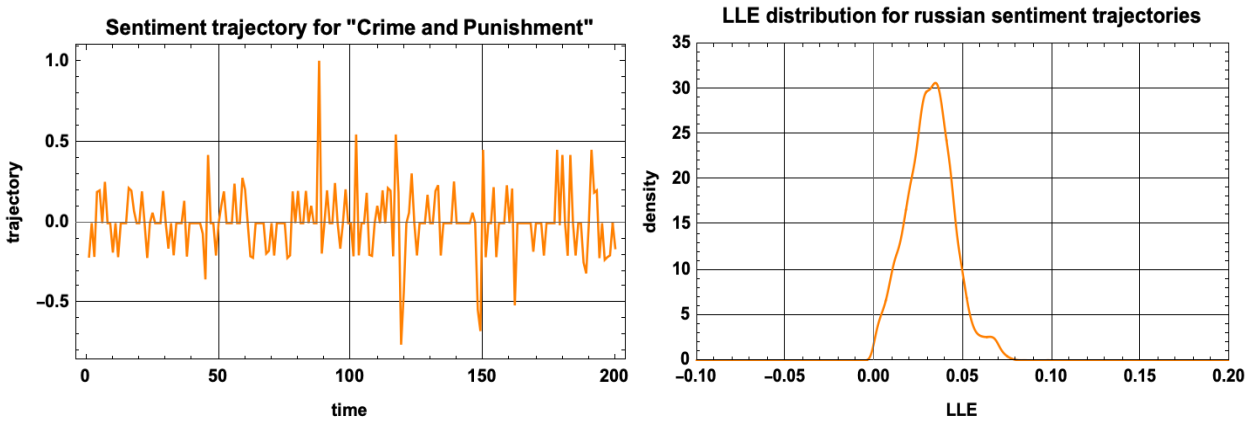


Fig 2 (left). Typical sentiment trajectory (F. Dostoyevsky's "Crime and Punishment")

Fig 3 (right). The LLE distribution for sentiment trajectories for Russian literature

So, the largest Lyapunov exponents are strictly positive for sentiment trajectories for all text, thereby indicating that the series are chaotic.

Figure 4a shows a distribution of entropy and complexity values over the respective plane for sentiment trajectories for the Russian literature; Figure 4b magnifies the appropriate part of it.

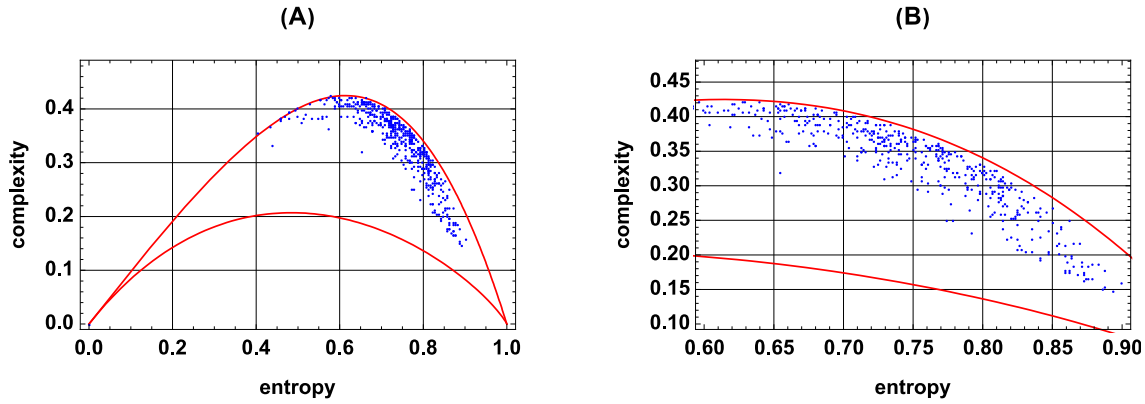


Fig 4. Entropy-complexity plane for sentiment trajectories for the Russian literature. Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

Minimum, average, and maximum entropy are equal to 0.13, 0.74, 0.96; they correspond, respectively, to "Like some giant from Sinai to Tabor …" by O. Mandelstam, "The Cave" by M. Tsvetaeva, "The Aesthetic Relations of Art to Reality" by N. Chernyshevsky. Minimum, average, and maximum complexity are equal to 0.10, 0.33, 0.42; they correspond, respectively, to N. Chernyshevsky's "The Aesthetic Relations of Art to Reality", S. Nadson's "In moments of sorrow, struggle and trial...", A. Maikov "Don't say that there is no escape...".

Evidently, points associated with most sentiment trajectories belong to the 'chaotic' area; some of them, to an area between chaos and noise (usually, it is a realm of financial time series).

Semantic trajectories

In view of the limitations of space, we discuss results in details for the Russian and English languages only; for other languages, we outline the results (also refer to Appendix).

Figure 5a shows a typical section of semantic trajectory for the Russian language (the first component; F. Dostoevsky, "Crime and Punishment"); Figure 5b, for the English language (first component; Ch. Dickens, "Oliver Twist"). For both figures, we use data corresponding to $d = 8, n = 1$.

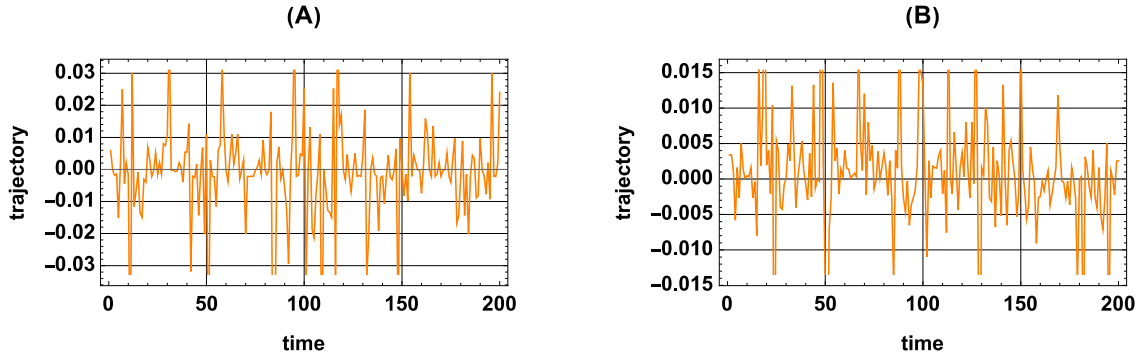


Fig 5. Typical semantic trajectory (the first component). (a) - F. Dostoyevsky's "Crime and Punishment", (b) – Ch. Dickens's "Oliver Twist".

Figure 6a demonstrates a distribution of the largest Lyapunov exponents for the Russian language; Figure 6b, for the English language (the first component). Evidently, the LLEs for both languages are strictly positive. This means that all sentiment trajectories are chaotic for both languages. For the Russian language, the minimum LLE is 8e-6 (it is achieved on V. Bryusov "The chill of morning spring..."); the maximum LLE is 7.4e-2 (it is achieved on V. Mayakovsky's "Let Us Take the New Rifles"); the average LLE is 3.05e-2 (it is achieved on "A Dog and a Horse" by I. Krylov). For the English language, the minimum LLE is 1.08e-3 (it is achieved on "The hardship upon the Ladies" by Jonathan Swift); the maximum LLE is 5.8e-2 (it is achieved on "The Tragical History of Dr. Faustus" by Christopher Marlowe); the average LLE is 4.2e-2 (it is achieved on "The Death Of The Old Year" by Lord Alfred Tennyson).

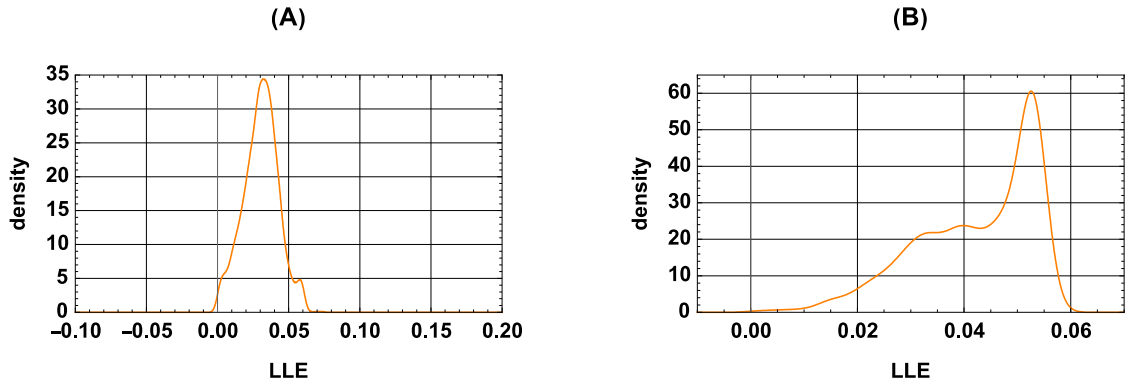


Fig 6. LLE distribution for semantic trajectories (n = 1, d = 1). (a) Russian literature; (b) English literature

Figure 7a exhibits distribution of the entropy-complexity pairs for the Russian literature; Fig 8a, for the English literature. Figures 7b and 8b magnify the respective areas of Figures 7a and 8a. All points of both Russian and English literature belong to the chaotic area. For the figures, we use data corresponding to $d = 4, n = 4$. S7-14 Figures in Appendix A present similar plots for other values of d, n (S1-14) and other languages (S15-S24). We can conclude (from the data) that for $d \geq 2, n \geq 3$ most semantic trajectories belong to the chaotic area.

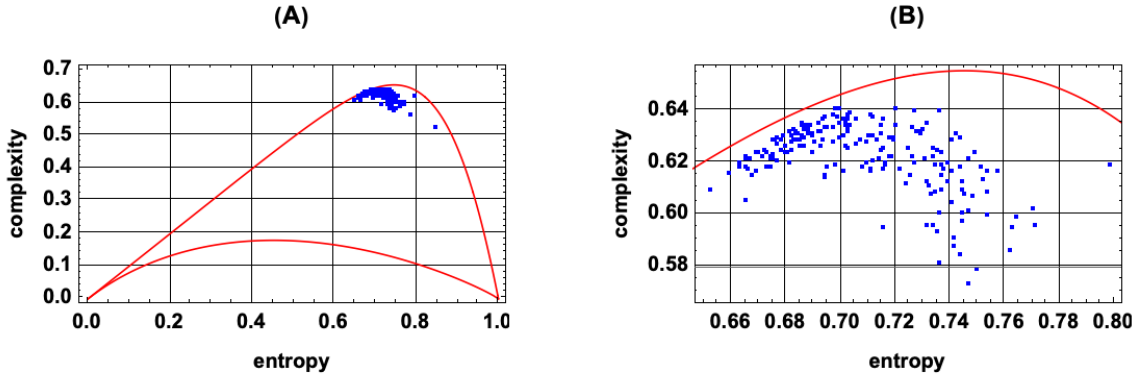


Fig 7. Entropy-complexity plane for semantic trajectories for the Russian literature. Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

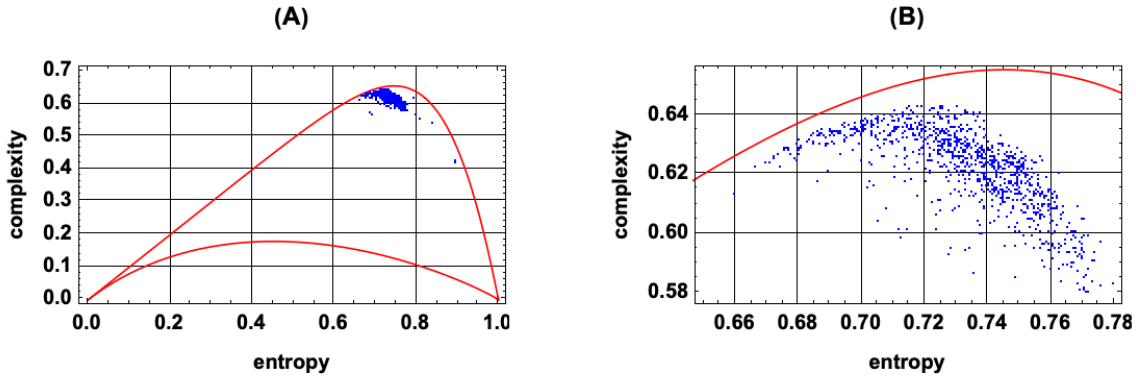


Fig 8. Entropy-complexity plane for semantic trajectories for the English literature. Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

We believe that lesser d's and n's, for which most texts appear to belong to the area of purely random processes, no longer reflect the nature of the texts. We can view these values of d and n as the lower boundary, such that semantic trajectories for larger d and n really reflect the respective text. On the other hand, as trajectories of literary masterpieces are limited in size, we are unable to estimate entropy and complexity adequately for very large values of d and n; this determines the higher boundary.

S2 Table in Appendix C lists minimum, average, and maximum values of entropy and complexity for the Russian literature (and titles of the respective texts); S3 Table (Appendix C), for the English.

Figure 9 shows typical dependence of the number of *FNN*s on the dimension of embedding space. The curve corresponds to "Anna Karenina", a classic novel by L. Tolstoy ($n = 5$). The plot demonstrates a typical 'chaotic' drop at $d = 5$.

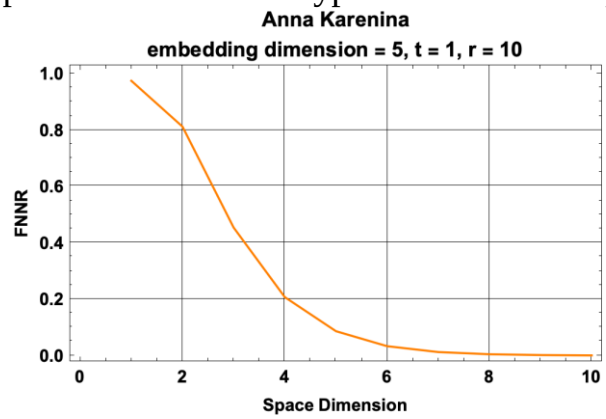


Fig 9. *FFN* rate dependence on the dimension of embedding space ("Anna Karenina" by L. Tolstoy ($n = 5$))

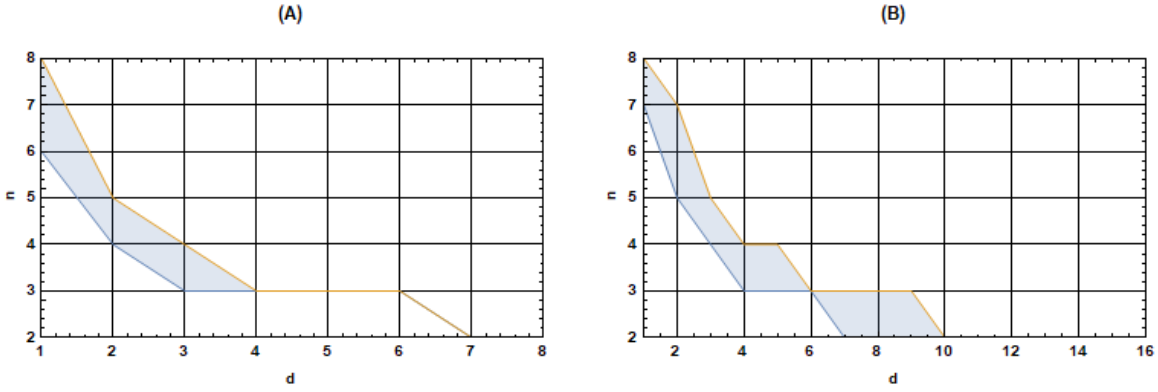


Fig 10. Admissible values of d, n for Russian (A) and English (B). Orange lines show borders for d and n, above which texts fall into the deterministic area; blue lines show borders, below which texts fall into the noise.

Figure 10 displays values of d, n such that one can analyse semantic trajectories of the Russian (Fig. 10a) and English (Fig. 10b) languages. In order to ascertain these values, we apply all three criteria discussed above. The blue area corresponds to admissible values of d and n; for values above the orange line texts fall into the deterministic area; for values below the blue line – the noise. Refer to Appendix B for similar figures for other languages.

For all languages considered, for large *d* (8 and more), the admissible values of *n* are small – we can consider bi-grams or tri-grams. For *d* = 1, the admissible values of *n* vary with a language. For example, for the French, German, and Norwegian languages one should consider only bi-grams, while for other languages the admissible values of n are 6 or higher. This fact affects the computational complexity of the algorithms in question.

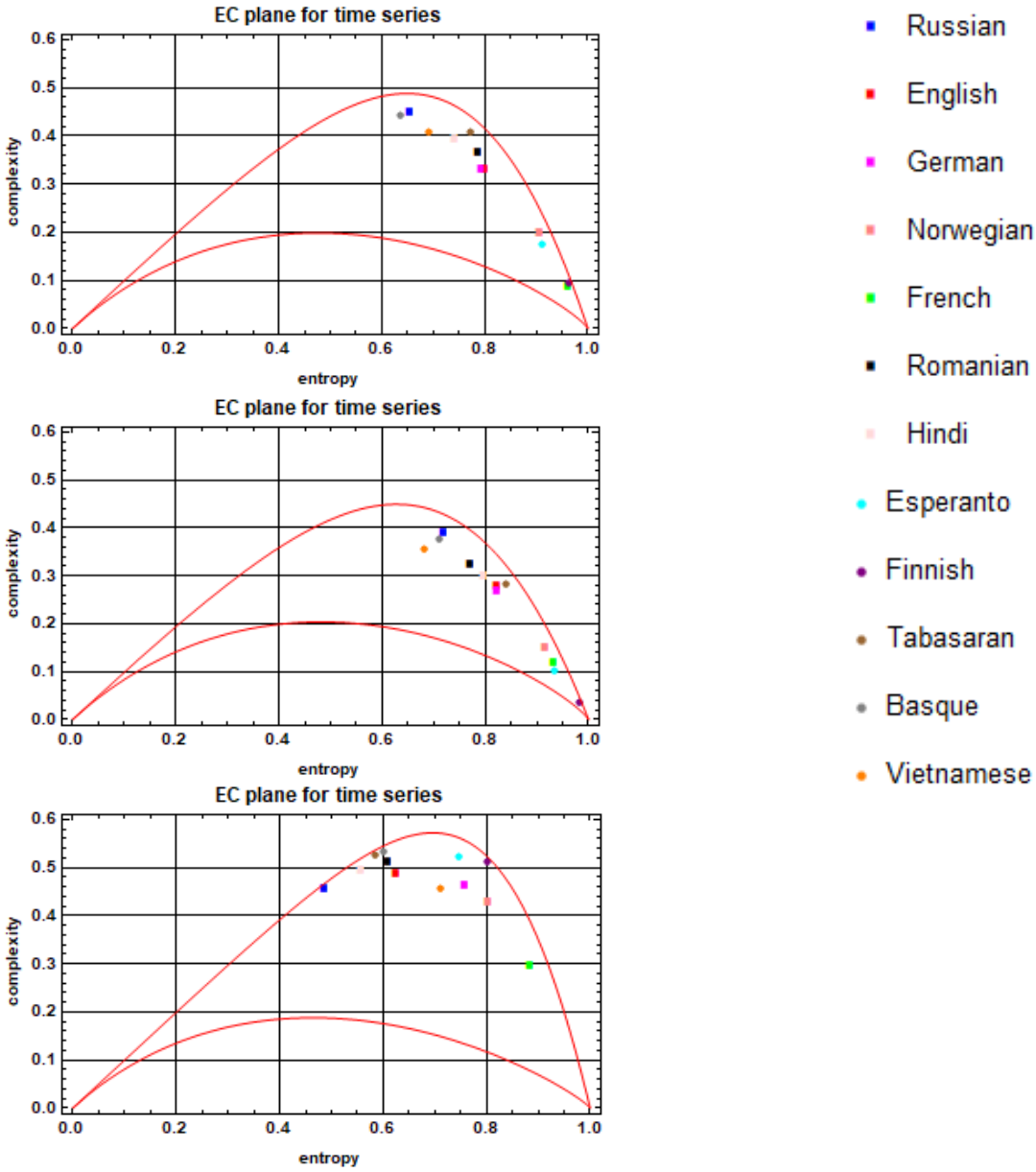


Fig 11. Mean entropy and complexity values for different languages.
d = 2, n = 4 (above); d = 3, n = 3 (centre); d = 5, n = 3 (below). Squared points mark Indo-European languages (see S1 Table in Appendix A).

Figure 11 shows mean values for entropy and complexity for various *d* and *n*. The Russian and Basque languages appear to be the most complex ones: the texts of these languages, on the average, closer to the upper boundary. The French language seems to be the closest to chaotic processes, most likely due to the big sample size of the French corpus (see S1 Table in Appendix A).

For various values of d and n, different languages cluster themselves into groups on the entropy-complexity plane. In general, these clusters coincide with language families. For $d = 3$, $n = 3$ Germanic and Italic languages of the Indo-European languages fall into one area, which can be seen in the central subfigure of Figure 11. The Esperanto language, for various values of d and n, 'approaches' to different Indo-European languages. For instance, for $d = 1$ it is close to the Romanian languages, for $d = 2$ it is close to Norwegian (see the first subfigure of Figure 11), for $d = 3$ it is close to French (central subfigure of Figure 11). Therefore, these results do not corroborate its claim to be equidistant from all languages. However, it seems to be equidistant from all Indo-European languages.

Trajectories of original masterpieces and their translations

These characteristics of semantic and sentient trajectories may be employed in order to assess the similarity in the style of an original masterpiece and its translation. Since the chaoticities of the languages differ, it is reasonable not to use absolute values of entropy and complexity, but the relative deviation from the average values for the respective language.

By way of illustration, we considered translations of "Oliver Twist" (by Ch. Dickens) into the Russian language. The novel has been translated several times: by A. Gorkovenko (XIX-th century), E. Lann and A. Krivtsova, and V. Lukianskaya (XX-th century). The corpus, which consists of classical masterpieces of Russian literature of the XX and XIX-th centuries, serves as a context for the translations; such a context seems to befit Dickens's translations. Besides that, we consider a machine word-by-word translation and a machine translation performed by a Facebook translator. S4 Table (Appendix D) summarises values of entropies and complexities for various d's and n's.

By way of illustration of Russian-to-English translation, we consider two classic novels: "Crime and Punishment" by F. Dostoevsky and "Anna Karenina" by L. Tolstoy. For both novels, we consider translations by K. Garnett, and R. Pevear and L. Volkhonskaya. As for the English-to-Russian translations, we also consider a machine word-by-word translation and a machine translation performed by a Facebook translator.

S5 and S6 Tables in Appendix D summarise values of entropies and complexities for these novels (more precisely, relative deviations thereof from the average value over the respective language, for various d's and n's. For "Crime and Punishment", the closest translation appears to be that by K. Garnett; for "Anna Karenina", that by K. Garnett. Again, machine translations appear to be farther from the original than human ones. Quite naturally, word-by-word translation appears to be the farthest.

Conclusions and future directions

Most literature masterpieces, for all languages considered, give rise to markedly chaotic sentiment and semantic trajectories. The respective 'entropy-complexity' pairs share 'chaotic' area with the conventional benchmarks of chaotic time series, similar to the Lorenz time series; the largest Lyapunov exponents are positive; dependence of the number of *FNN*s on the embedding dimension demonstrates rapid drop, as it should be for a chaotic series; the value at which this happens is a rough estimate of the strange attractor dimension. The Russian language turns out to be more 'chaotic' than, for example, the English one; we attribute this fact to the free order of words. Russian is characterized by flexible word order in a sentence, as opposed to English.

We estimated high and low boundaries for the size of n-grams and embedding dimensions, which guarantee that the semantic trajectories reflect the true dynamics of the respective literature masterpiece in a semantic space (because of the Takens's theorem).

For various values of d and n, different languages cluster themselves into groups on the entropy-complexity plane. In general, these clusters coincide with language families. The Esperanto language, for various values of d and n, 'approaches' to different Indo-European languages. The results do not corroborate its claim to be equidistant from all languages. However, it seems to be equidistant from all Indo-European languages.

The results for each literature masterpiece are available on request.

We employ this approach to compare styles of an original masterpiece and its translations, both human and machine. It appears that machine translations are still worse than human ones, however, for example, the Facebook translation is comparable with them.

As future directions, we propose to study the difference between texts written by humans and those generated by bots with the employment of the characteristics of sentiment and semantic trajectories. In addition, it seems interesting to explore other languages, in the framework both conventional language typology (based upon genetic relationship of the languages) [58] and language typology, based upon principles and parameters theory [59].

Acknowledgements

The author is deeply indebted to Mr J. Cumberland, HSE University for proof-editing. This research was supported in part through computational resources of HPC facilities at HSE University.

References

1. Reagan AJ, Mitchell L, Kiley D, Danforth CM, Dodds PS. The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science. 2016; 5(1): 1.
2. Wang H, Wu H, He Z, Huang L, Church KW. Progress in Machine Translation. Engineering. 2021.
3. Tan Z, Wang S, Yang Z, Chen G, Huang X, Sun M, et al. Neural machine translation: A review of methods, resources, and tools. AI Open. 2020; 1: 5–21.
4. Offord D. Using Russian: A Guide to Contemporary Usage. Cambridge University Press; 1996.
5. Shopen T. Language Typology and Syntactic Description. Grammatical Categories and the Lexicon, Cambridge University Press. 2007; 3.
6. Català N, Baixeries J, Ferrer-i-Cancho R, Padró L, Hernández-Fernández A Zipf's laws of meaning in Catalan. PLoS ONE. 2021; 16(12): e0260849.
7. Cancho RFI, Solé RV. The small world of human language. Proceedings of the Royal Society of London. 2001; Series B: Biological Sciences, 268(1482): 2261-2265.
8. Barrát A, Barthélemy M, Vespignani A. Dynamical Processes on Complex Networks May. Journal of Statistical Physics. 2009; 135(4): 773-774.
9. Newman M. Networks: An Introduction. Oxford University Press; 2010.
10. Elson K, Dames N, McKeown KR. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics; 2010. Association for Computational Linguistics; 2010. p. 138-147.
11. Mac Carron P, Kenna R. Network analysis of the Íslendinga sögur–the Sagas of Icelanders. The European Physical Journal B 2013; 86(10): 1-9.
12. Kydros D, Notopoulos P, Exarchos E. Homer's Iliad: A Social Network Analytic Approach. International Journal of Humanities and Arts Computing. 2015; 9: 115-132.
13. Waumans MC, Nicodème T, Bersini H. Topology analysis of social networks extracted from literature. PLoS One. 2015; 10(6): e0126470.
14. Stella M, Brede M. Mental lexicon growth modelling reveals the multiplexity of the English language. In: Complex Networks VII. Springer, Cham; 2016. p. 267-279.
15. Dodds PS, Clark EM, Desu S, Frank MR, Reagan AJ, Williams JR, et al. Human language reveals a universal positivity bias. Proceedings of the national academy of sciences. 2015; 112(8): 2389-2394.
16. Min S, Park J. Mapping out narrative structures and dynamics using networks and textual information. arXiv:1604.03029 [Preprint]. 2016 [cited 2020 Jan 16]: [17 p]. Available from: https://arxiv.org/abs/1604.03029
17. Kiley DP, Reagan AJ, Mitchell L, Danforth CM, Dodds PS. Game story space of professional sports: Australian rules football. Physical Review E. 2016; 93(5), 052314.
18. Boucher J, Osgood CE. The Pollyanna hypothesis. Journal of verbal learning and verbal behavior. 1969; 8(1): 1-8.

19. Min S, Park J, Network Science and Narratives: Basic Model and Application to Victor Hugo's Les Misérables. In: Complex Networks VII. Springer. p. 257-265.
20. Debowski L. Information Theory Meets Power Laws: Stochastic Processes and Language Models. John Wiley & Sons; 2020.
21. Tanaka-Ishii K. Statistical Universals of Language. Springer; 2021.
22. Gromov VA, Migrina AM. A Language as a Self-Organized Critical System. Complexity. 2017. ArticleID 9212538.
23. Scheffer M, van de Leemput I, Weinans E, Bollen J. The rise and fall of rationality in language. Proceedings of the National Academy of Sciences. 2021; 118(51).
24. Doherty S. The Impact of Translation Technologies on the Process and Product of Translation. International Journal of Communication. 2016; 10: 947-969.
25. House J. Translation Quality Assessment. Past and present. Routledge; 2015.
26. Koehn P. Statistical machine translation. Cambridge University Press; 2009.
27. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; 2002; Philadelphia, Pennsylvania, USA. 311-318.
28. Liu X, Zhao J, Sun S, Liu H, Yang H. Variational multimodal machine translation with underlying semantic alignment. Information Fusion 69. 2021; 73–80.
29. Munk M, Munkova D, Benko L. Towards the use of entropy as a measure for the reliability of automatic MT evaluation metrics. Journal of Intelligent & Fuzzy Systems 34. 2018; 3225–3233.
30. Munkova D, Hajek P, Munka M, Skalka J. Evaluation of Machine Translation Quality through the Metrics of Error Rates and Accuracy. Procedia Computer Science 171. 2020; 1327–1336.
31. Esplà-Gomis M, Sánchez-Martínez F, Mikel L. Forcada Predicting insertion positions in word-level machine translation quality estimation. Applied Soft Computing Journal 76. 2019; 174–192.
32. Luong NQ, Besacier L, Lecouteux B. Towards accurate predictors of word quality for Machine Translation: Lessons learned on French–English and English–Spanish systems. Data & Knowledge Engineering. 2015; 96: 32-42.
33. Le QV, Schuster M. A neural network for machine translation, at production scale. [Internet]. 2016 Sep 27 [cited 2020 Jan 16]. Available from: https://research.googleblog.com/2016/09/a-neural-network-for-machine.html
34. Jia Y, Carl M, Wang X. How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. The Journal of Specialised Translation 31. 2019; 61-86.
35. Chon YV, Shin D, Kim GE. Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. System 96. 2021; 102408.

36. Nguyen T, Nguyen L, Tran P, Nguyen H. Improving Transformer-Based Neural Machine Translation with Prior Alignments. Complexity. 2021. Article ID 5515407.
37. Jarvis S. Defining and measuring lexical diversity. Vocabulary knowledge: Human ratings and automated measures. 2013: 13-45.
38. Graesser AC, McNamara DS, Kulikowich JM. Coh-Metrix: Providing multilevel analyses of text characteristics. Educational Researcher. 2011; 40(5):223-234.
39. McCarthy PM, Jarvis S. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods. 2010; 42(2): 381-392.
40. Wang H, Wu H, He Z, Huang L, Church KW. Progress in Machine Translation. Engineering. 2021.
41. Bellegarda JR. Latent Semantic Mapping: Principles & Applications. Morgan & Claypool; 2007.
42. Golub G, Kahan W. Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics. 1965; Series B: Numerical Analysis, 2(2): 205-224.
43. Kalman D. A singularly valuable decomposition: the SVD of a matrix. The college mathematics journal 1996; 27(1): 2-23.
44. Balleza E, Alvarez-Buylla ER, Chaos A, Kauffman S, Shmulevich I, Aldana M. Critical Dynamics in Genetic Regulatory Networks: Examples from Four Kingdoms. PLoS ONE. 2008; 3(6): e2456.
45. Mangin L, Lesèche G, Duprey A, Clerici C. Ventilatory Chaos Is Impaired in Carotid Atherosclerosis. PLoS ONE. 2011; 6(1): e16297.
46. Shapoval A. Prediction problem for target events based on the inter-event waiting time. Physica A: Statistical Mechanics and its Applications. 2010; 389(22): 5145-5154.
47. Shapoval A, Le Mouël J-L, Courtillot V, Shnirman M. Two regimes in the regularity of sunspot number. The Astrophysical Journal. 2013; 779(2): 108.
48. Gromov VA, Borisenko EA. Chaotic time series prediction and clustering methods, Neural Computing and Applications. 2015; 2: 307-315.
49. Gromov VA, Konev AS. Precocious identification of popular topics on Twitter with the employment of predictive clustering. Neural Computing and Applications. 2017; 28(11): 3317–3322.
50. Rosso OA, Larrondo HA, Martin MT, Plastino A, Fuentes MA. Distinguishing noise from chaos. Physical review letters. 2007; 99(15), 154102.
51. Kantz H, Schreiber T. Nonlinear time series analysis. Cambridge university press. 2004; 7.
52. Malinetsky GG, Potapov AB. [Current Problems in Nonlinear Dynamics] Sovremennye problemy nelineinoi dinamiki. Moscow: Editorial URSS; 2000. Russian.
53. Bandt C, Pompe B. Permutation entropy: a natural complexity measure for time series. Physical review letters. 2002; 88(17), 174102.

54. Rosenstein MT, Collins JJ, De Luca CJ. A practical method for calculating largest Lyapunov exponents from small data sets. Physica D: Nonlinear Phenomena. 1993; 65(1-2): 117-134.
55. Furstenberg H. Poincaré recurrence and number theory. Bulletin (New Series) of the American Mathematical Society. 1981; 5(3): 211-234.
56. Hegger R, Kantz H, Schreiber T. Practical implementation of nonlinear time series methods: The TISEAN package. Chaos: An Interdisciplinary Journal of Nonlinear Science. 1999; 9(2): 413-435.
57. Takens F. Detecting strange attractors in turbulence. In Dynamical systems and turbulence, Warwick 1980. Berlin, Heidelberg: Springer; 1981. p. 366-381.
58. Murray G.-M., Peiros I., Starostin G. Distant language relationship: The current perspective. Journal of Language Relationship. 2009; 5: 13-30.
59. Newmeyer F. J. Possible and probable languages: A generative perspective on linguistic typology. Oxford University Press on Demand; 2005.

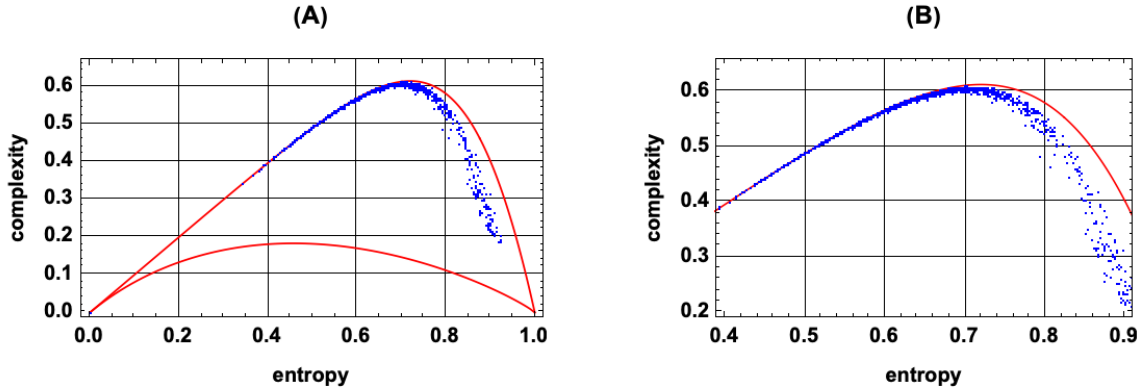
Supporting Information

Appendix A.

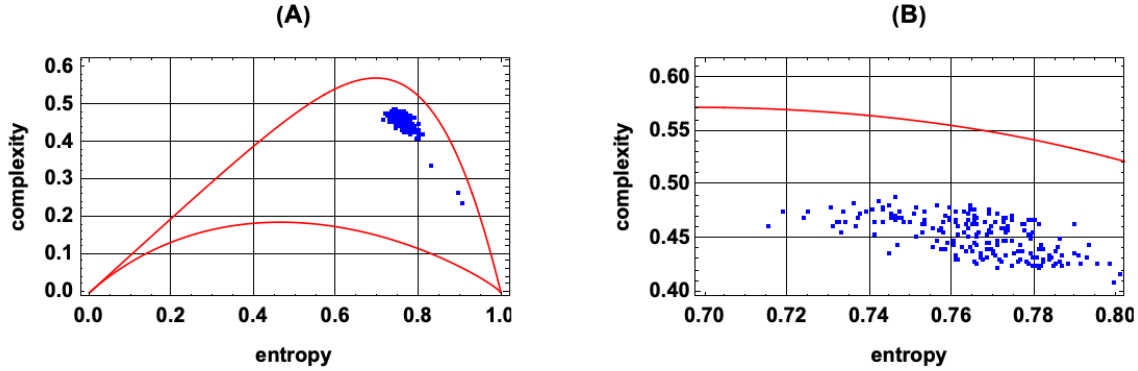
S1 Table. Language corpora details.

Language	Language family / group	Corpus size	Sources	Library
Russian	Indo-European / East Slavic	12683	Literature	natasha
English	Indo-European / Germanic	11052	Literature	spacy
German	Indo-European / Germanic	12503	Literature	spacy
Norwegian	Indo-European / Germanic	4124	Literature	spacy
French	Indo-European / Italic	1568	Literature	spacy
Romanian	Indo-European / Italic	2374	Literature	spacy
Hindi	Indo-European / Indo-Iranian	1043	Literature	StanfordNLP
Finnish	Uralic / Finno-Ugric	3385	Literature	uralicNLP
Tabasaran	Northeast Caucasian / Lezgic	1386	News	manual preprocessing
Esperanto	Constructed language	1171	Literature	esperanto-analyzer
Basque	Language isolate	10052	Wikipedia	SparkNLP
Vietnamese	Austroasiactic / Vietic	1071	Literature	pyvi

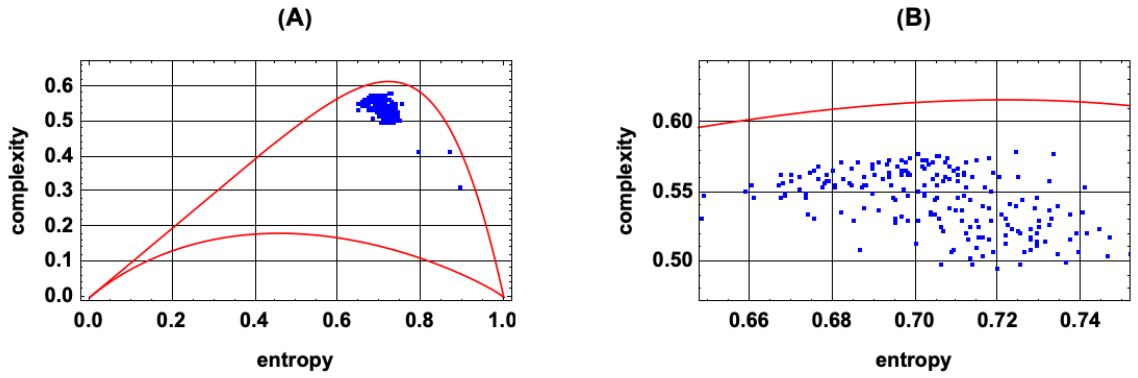
Distributions of entropy-complexity pairs for semantic trajectories for various *d*'s and *n*'s.



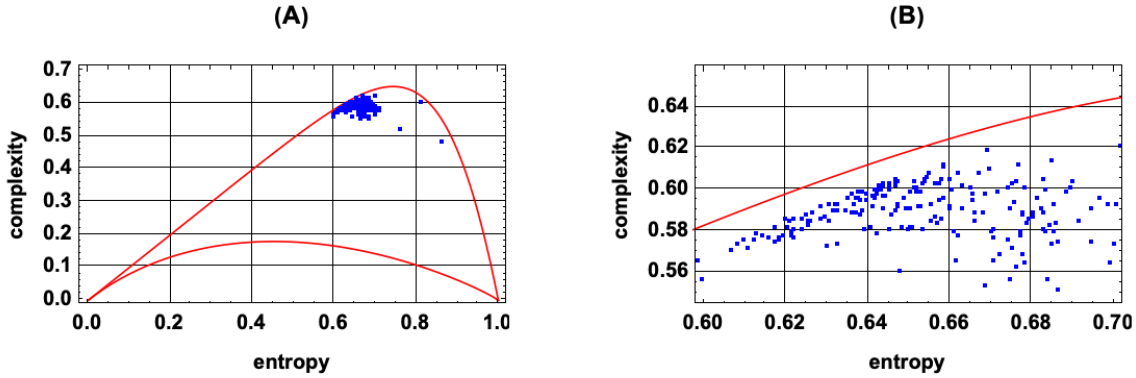
S1 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=1, n=8). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



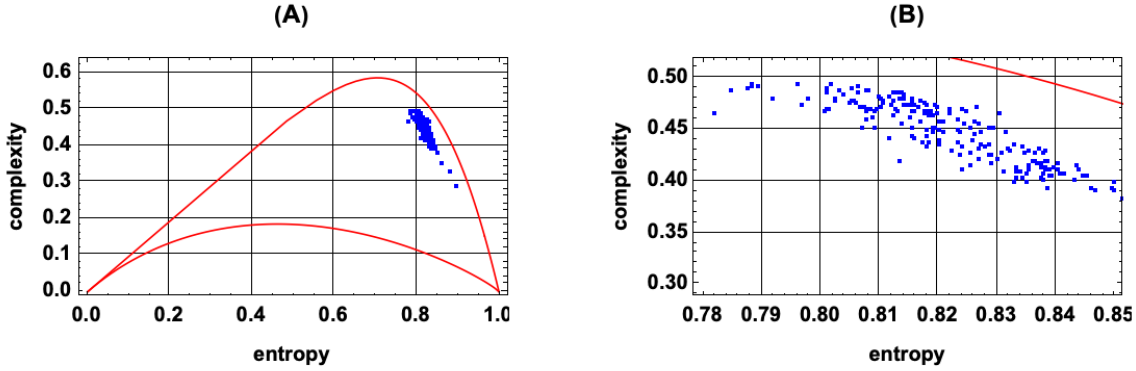
S2 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=5, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



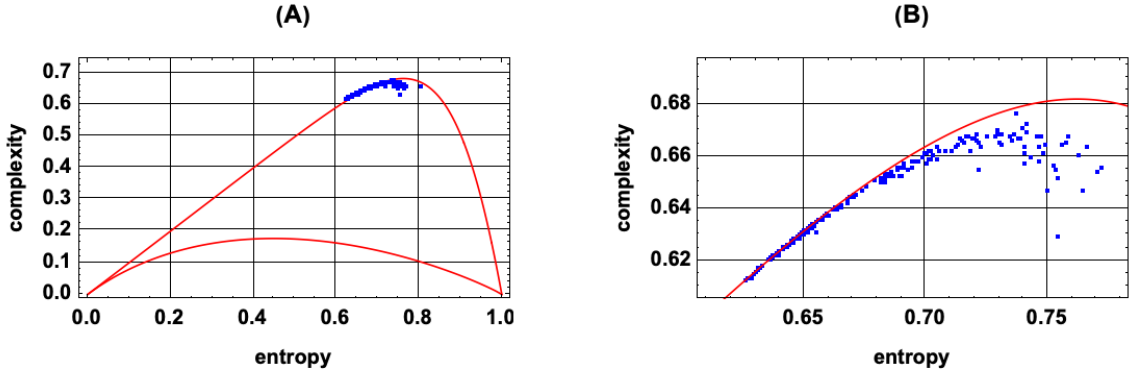
S3 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=6, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



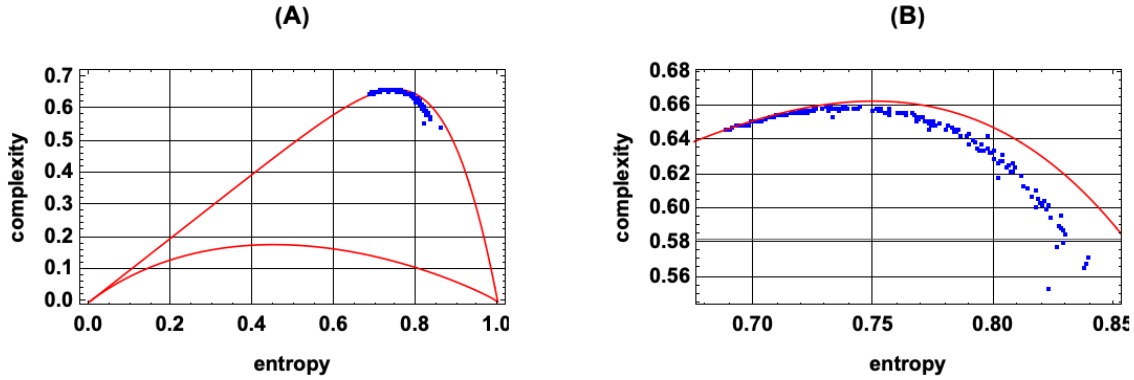
S4 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=7, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



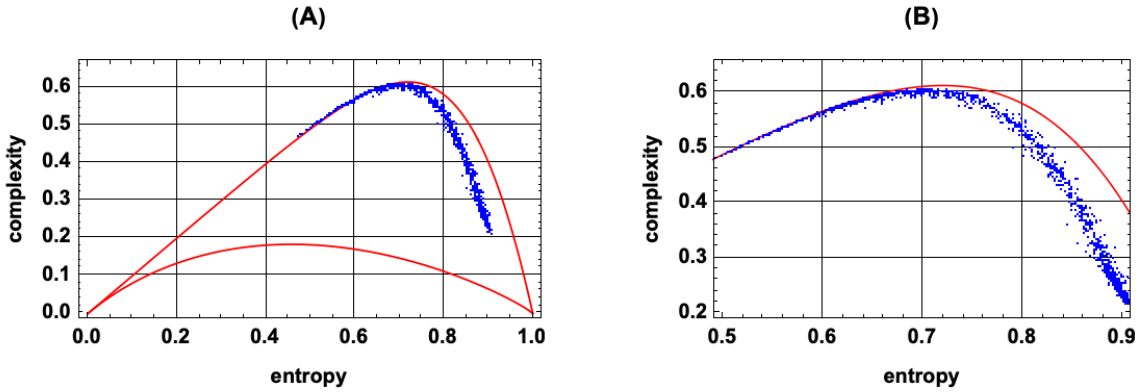
S5 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=3, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



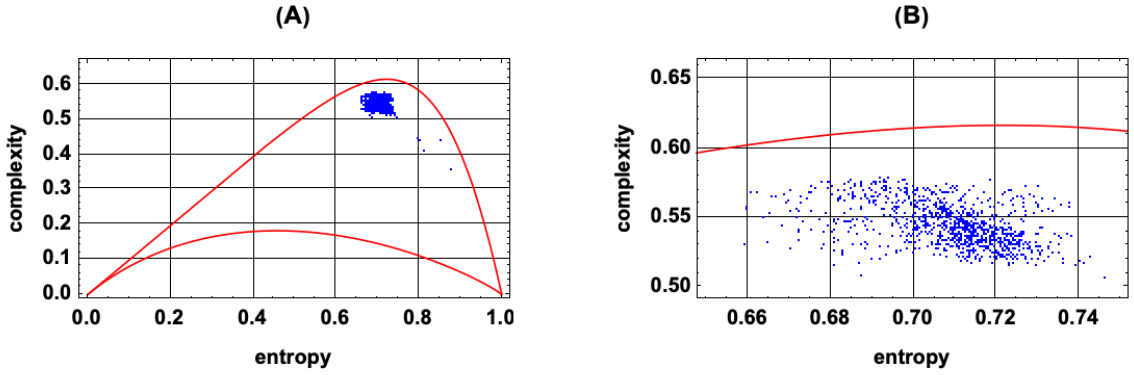
S6 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=3, n=5). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



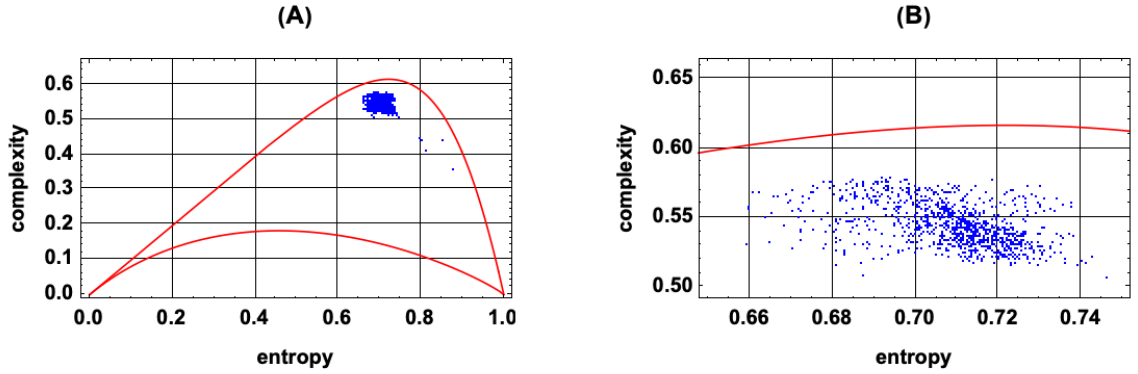
S7 Fig. Entropy-complexity plane for semantic trajectories for the Russian literature for (d=2, n=6). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



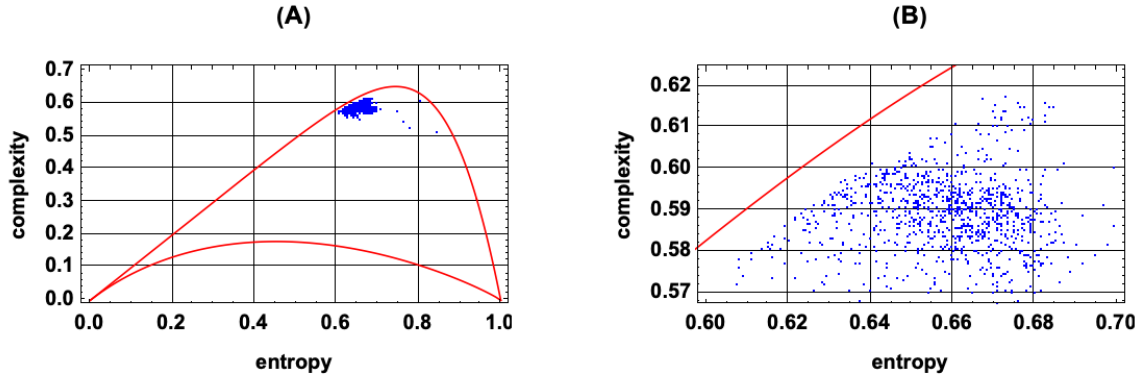
S8 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=1, n=8). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



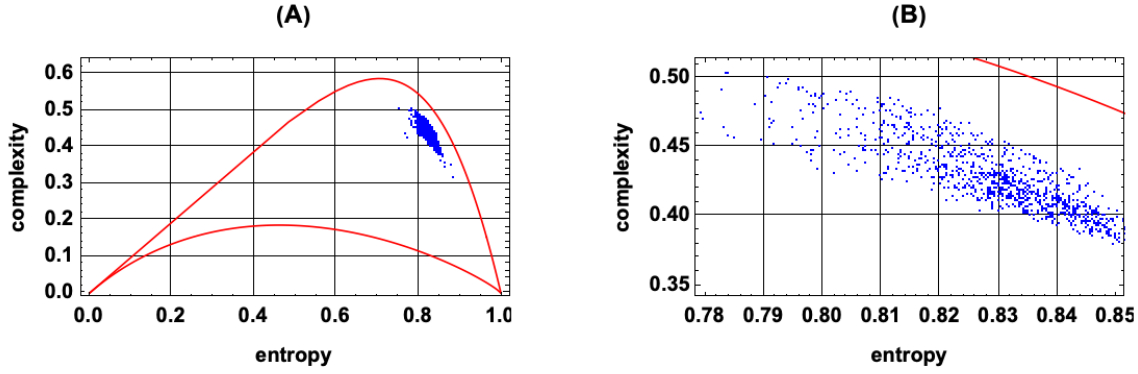
S9 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=5, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



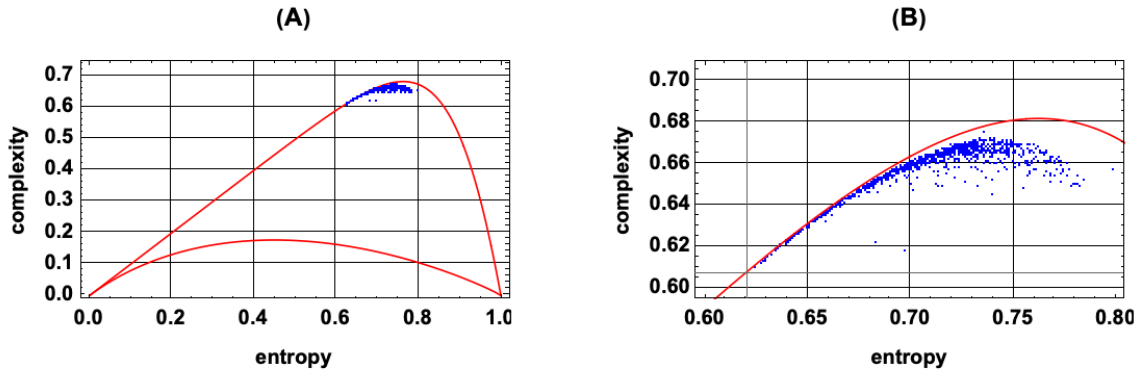
S10 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=6, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



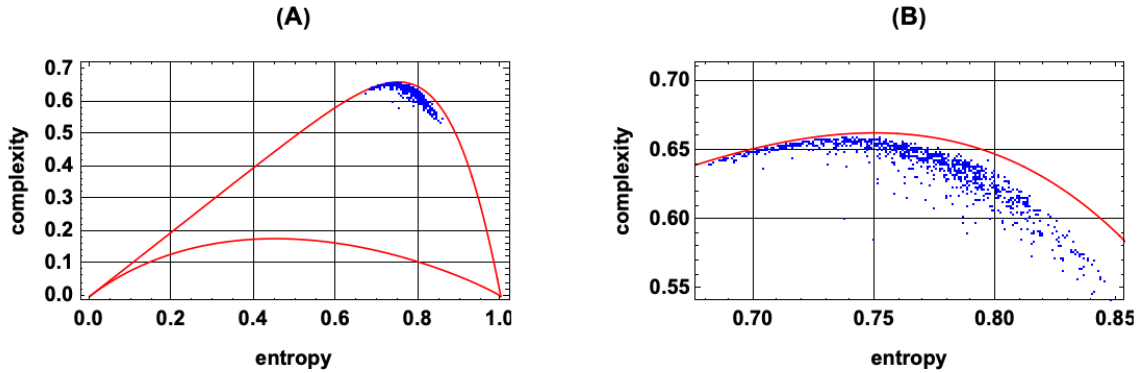
S11 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=7, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



S12 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=3, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

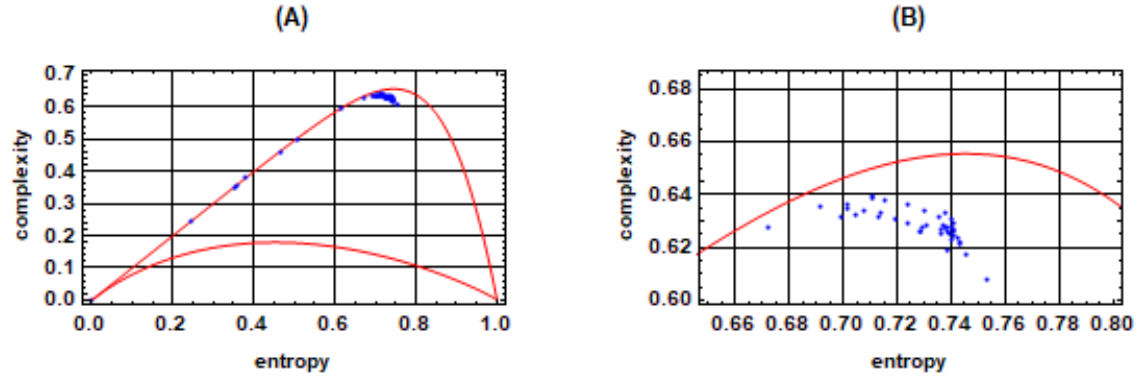


S13 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=3, n=5). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

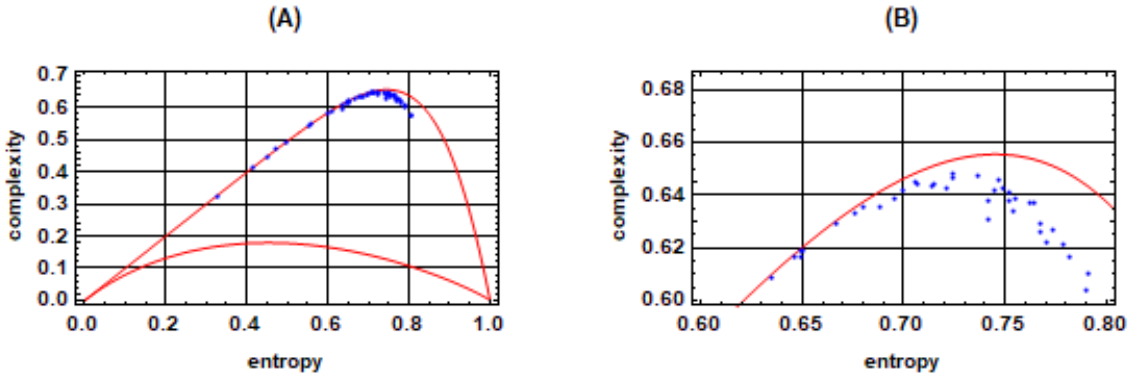


S14 Fig. Entropy-complexity plane for semantic trajectories for the English literature for (d=2, n=6). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

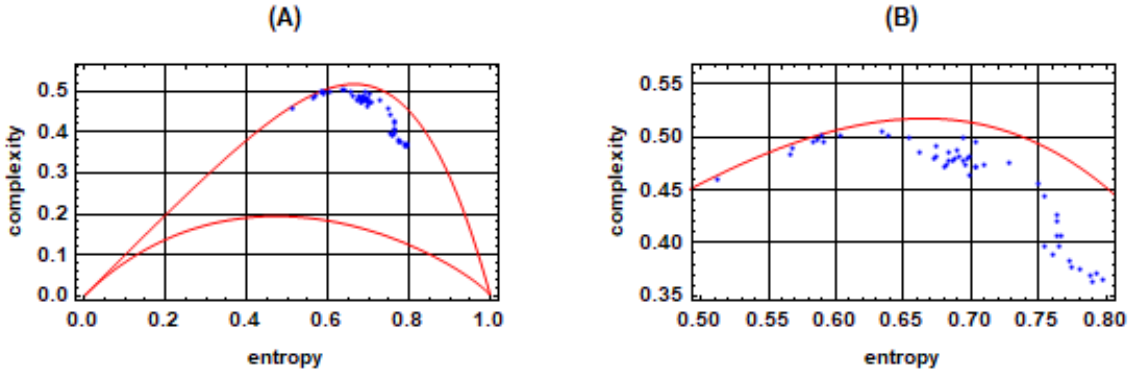
Distributions of entropy-complexity pairs for semantic trajectories for various languages.



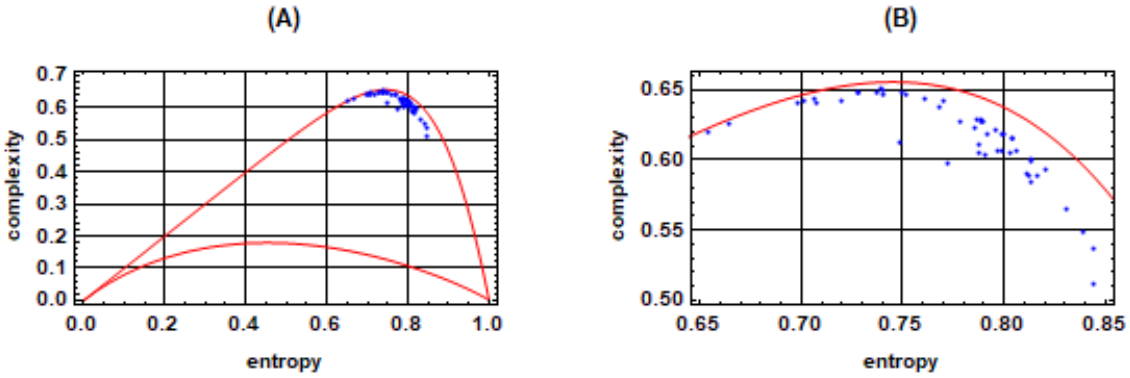
S15 Fig. Entropy-complexity plane for semantic trajectories for the German literature for (d=4, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



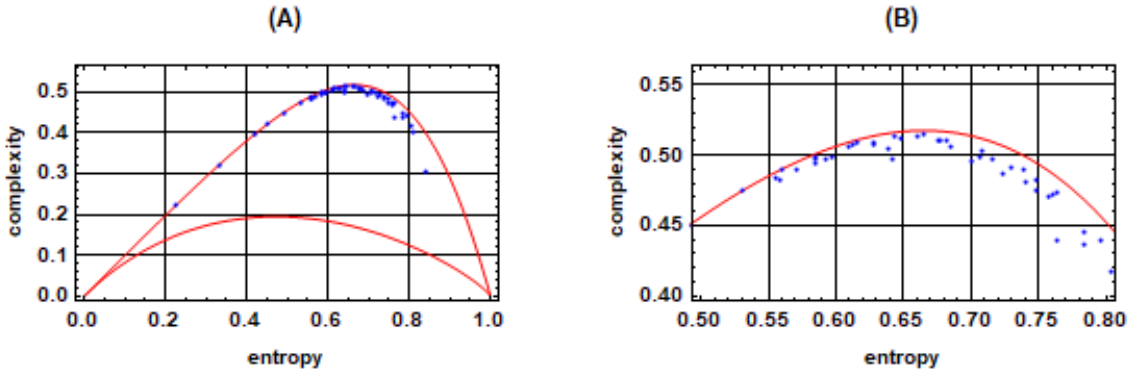
S16 Fig. Entropy-complexity plane for semantic trajectories for the Norwegian literature for (d=4, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



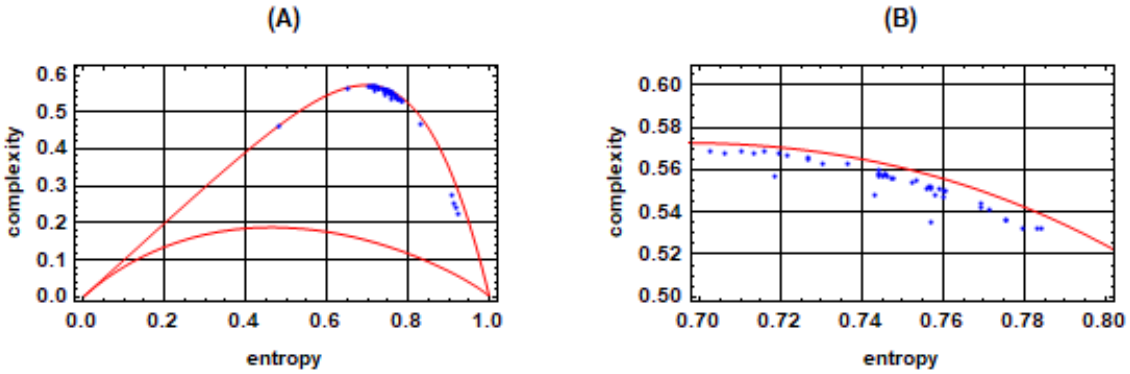
S17 Fig. Entropy-complexity plane for semantic trajectories for the Romanian literature for (d=4, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



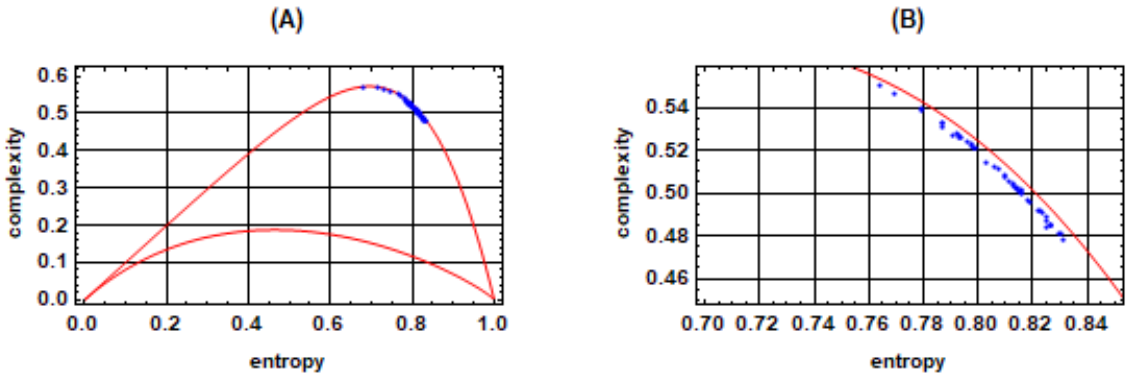
S18 Fig. Entropy-complexity plane for semantic trajectories for the French literature for (d=4, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



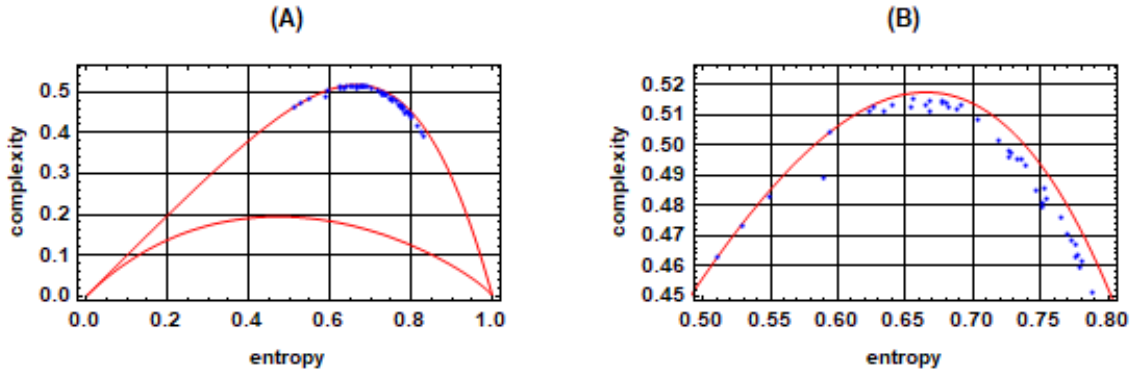
S19 Fig. Entropy-complexity plane for semantic trajectories for the Hindi literature for (d=4, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



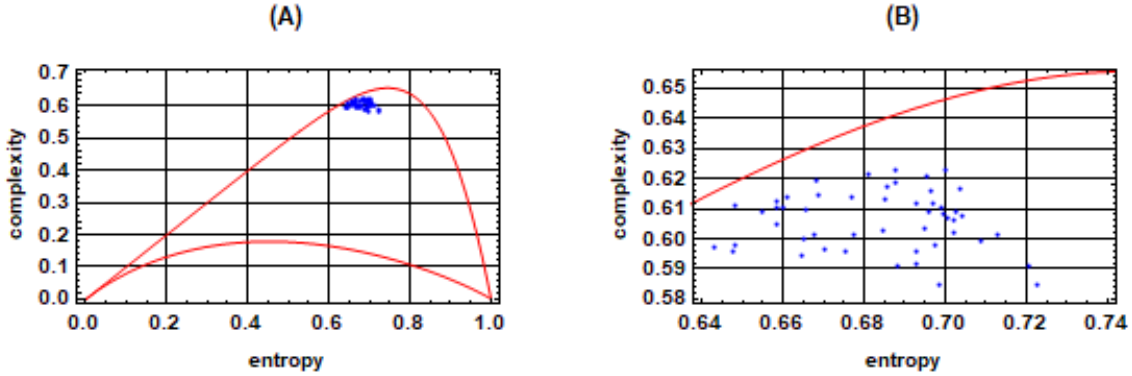
S20 Fig. Entropy-complexity plane for semantic trajectories for the Esperanto literature for (d=5, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



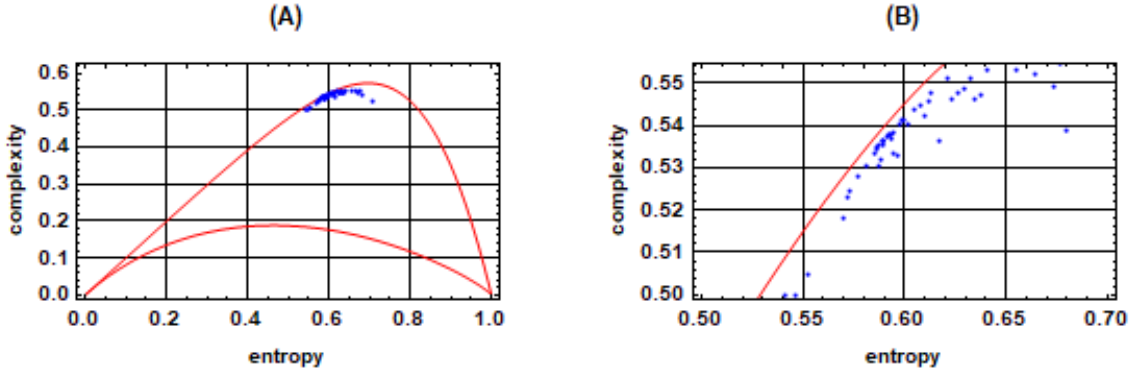
S21 Fig. Entropy-complexity plane for semantic trajectories for the Finnish literature for (d=5, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.



S22 Fig. Entropy-complexity plane for semantic trajectories for the Tabasaran literature for (d=4, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

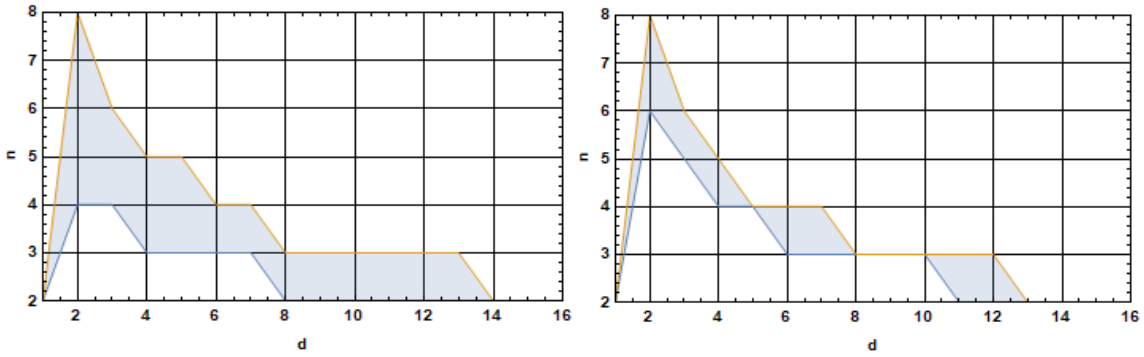


S23 Fig. Entropy-complexity plane for semantic trajectories for the Vietnamese literature for (d=4, n=4). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

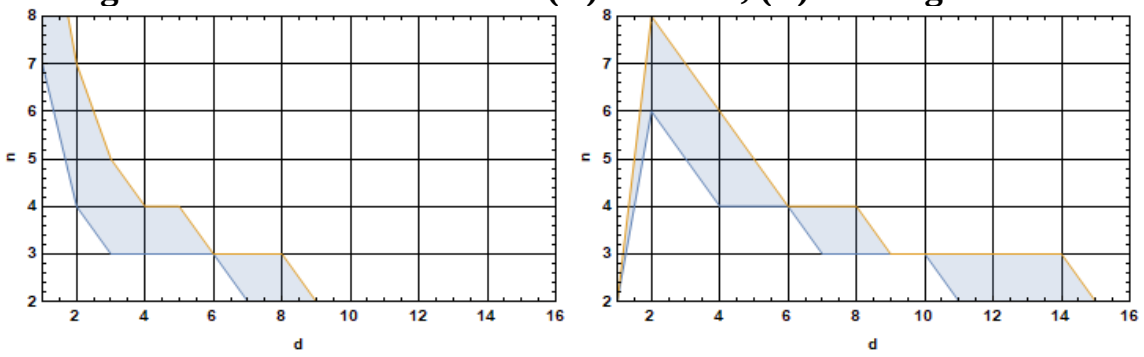


S24 Fig. Entropy-complexity plane for semantic trajectories for the Basque literature for (d=5, n=3). Red solid curves show theoretical boundaries, blue dots show sentiment trajectories. (A) full plane, (B) magnified part of the plane.

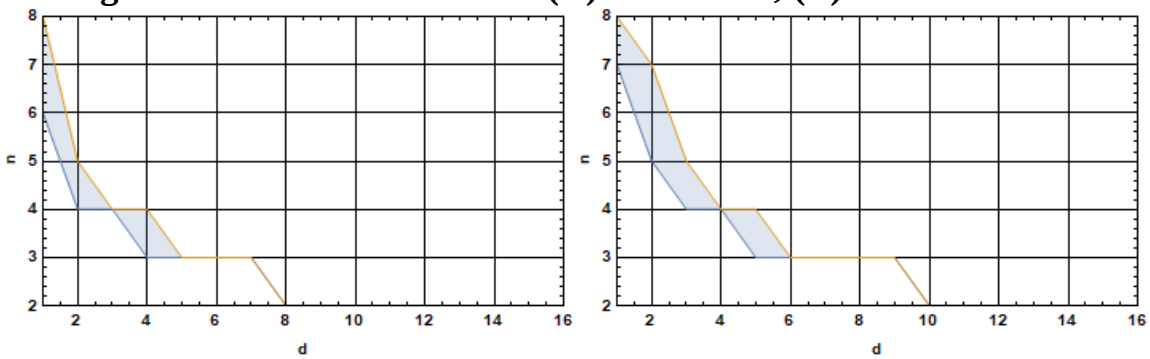
Appendix B. Admissible values of *d* and *n* for different languages.



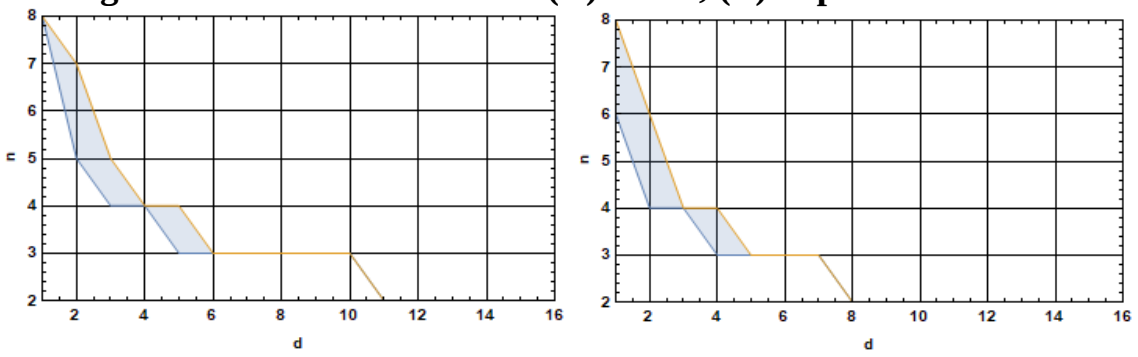
S25 Fig. Admissible values for the (A) German, (B) Norwegian Literature



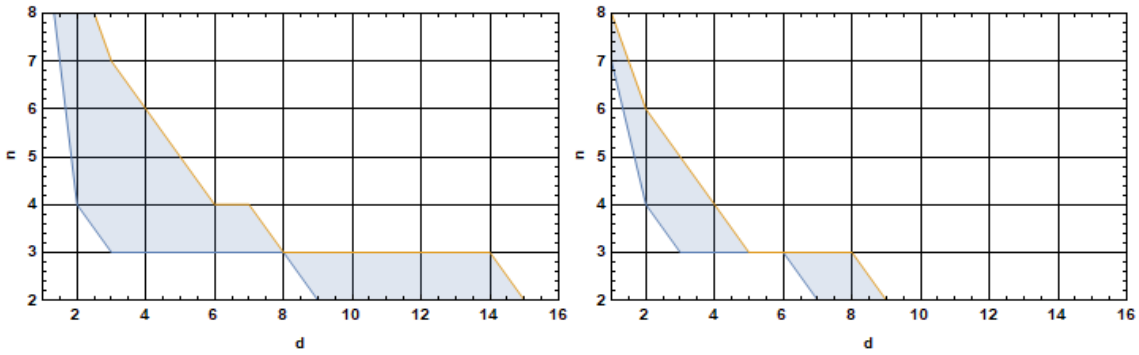
S26 Fig. Admissible values for the (A) Romanian, (B) French Literature



S27 Fig. Admissible values for the (A) Hindi, (B) Esperanto Literature



S28 Fig. Admissible values for the (A) Finnish, (B) Tabasaran Literature



S29 Fig. Admissible values for the (A) Vietnamese, (B) Basque Literature

Appendix C. Minimum, average, and maximum entropy and complexity values.

S2 Table. Minimum, average, and maximum entropies and complexities for texts of the Russian language

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
3	5	Minimum entropy	0.715	0.461	-6.68%	1.81%	I. Turgenev "A Month in the Country"
		Mean entropy	0.766	0.426	-0.02%	-0.36%	A. Pushkin, "The Captain's Daughter"
			0.766	0.426	-0.02%	-5.94%	I. Goncharov "The Cliff"
		Maximum entropy	0.813	0.421	6.06%	-6.85%	S. Bobrov, "Taurica, or My Summer Day at the Taurian Chersonesus"
		Minimum complexity	0.799	0.408	4.28%	-9.79%	N. Karamzin, "Letters of a Russian Traveller"
		Mean complexity	0.779	0.452	1.59%	-0.02%	A. Bestuzhev, "The Test"
		Maximum complexity	0.746	0.489	-2.63%	8.04%	A. Remizov, "On Free Bread"
		Mean entropy and complexity	0.766	0.453	-0.04%	0.10%	L. Andreev, "Anathema"
3	6	Minimum entropy	0.649	0.530	-8.01%	-2.66%	I. Turgenev "A Month in the Country"
		Mean entropy	0.705	0.534	-0.01%	-2.03%	I. Turgenev "A Nest of the Gentlefolk"
		Maximum entropy	0.755	0.551	7.02%	1.18%	S. Bobrov, "Taurica, or My Summer Day at the Taurian Chersonesus"
		Minimum complexity	0.720	0.495	2.15%	-9.12%	F. Dostoevsky, "The Brothers Karamazov"
		Mean complexity	0.713	0.545	1.18%	-0.05%	V. Rozanov, "The First Box"

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
			0.705	0.545	-0.05%	-0.05%	M. Bulgakov, “A Young Doctor’s Notebook”
		Maximum complexity	0.724	0.579	2.74%	6.31%	M. Tsvetaeva, “The Maiden Tsar”
		Mean entropy and complexity	0.705	0.545	-0.05%	-0.05%	M. Bulgakov, “A Young Doctor’s Notebook”
3	7	Minimum entropy	0.599	0.565	-8.49%	-4.13%	L. Tolstoy, “The Living Corpse”
		Mean entropy	0.654	0.604	0.00%	2.48%	I. Bunin, “Mitya’s Love”
		Maximum entropy	0.712	0.575	8.91%	-2.35%	D. Merezhkovsky, “Leonardo da Vinci: Gods Resurgent”
		Minimum complexity	0.687	0.552	4.95%	-6.38%	L. Tolstoy, “War and Peace”
		Mean complexity	0.633	0.589	-3.24%	0.01%	A. Chekhov, “Three Sisters”
		Maximum complexity	0.702	0.621	7.29%	5.39%	S. Bobrov, “Taurica, or My Summer Day at the Taurian Chersonesus”
		Mean entropy and complexity	0.653	0.591	-0.21%	0.36%	V. Veresaev, “At the Turn”
4	3	Minimum entropy	0.782	0.465	-4.77%	3.70%	I. Turgenev “A Month in the Country”
		Mean entropy	0.821	0.447	-0.04%	-0.18%	M. Bulgakov, “A Young Doctor’s Notebook”
		Maximum entropy	0.852	0.382	3.67%	-14.82%	N. Karamzin, “Letters of a Russian Traveller”
		Minimum complexity	0.852	0.382	3.67%	-14.82%	N. Karamzin, “Letters of a Russian Traveller”
		Mean complexity	0.818	0.448	-0.39%	0.00%	L. Tolstoy, “Childhood”
		Maximum complexity	0.788	0.494	-4.03%	10.14%	M. Kheraskov, “The Hater”
		Mean entropy and complexity	0.821	0.447	-0.04%	-0.18%	M. Bulgakov, “A Young Doctor’s Notebook”
4	4	Minimum entropy	0.653	0.609	-7.63%	-2.18%	L. Tolstoy, “The Living Corpse”
		Mean entropy	0.708	0.628	0.12%	0.88%	M. Bulgakov, “A Young Doctor’s Notebook”

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
		Maximum entropy	0.771	0.596	9.08%	-4.36%	D. Merezhkovsky, “Leonardo da Vinci: Gods Resurgent”
		Minimum complexity	0.747	0.573	5.64%	-8.07%	L. Tolstoy, “War and Peace”
		Mean complexity	0.676	0.623	-4.45%	0.04%	A. Pisemsky, “A Bitter Fate”
		Maximum complexity	0.700	0.641	-1.03%	2.89%	A. Pushkin, “Ruslan and Ludmila”
		Mean entropy and complexity	0.706	0.627	-0.20%	0.69%	A. Pushkin, “The Captain’s Daughter”
5	3	Minimum entropy	0.626	0.612	-8.24%	-5.07%	M. Kheraskov, “The Hater”
		Mean entropy	0.683	0.651	0.02%	0.92%	A. Chekhov, “The Duel”
		Maximum entropy	0.771	0.654	12.94%	1.46%	I. Goncharov, “The Frigate Pallada”
		Minimum complexity	0.626	0.612	-8.24%	-5.07%	M. Kheraskov, “The Hater”
		Mean complexity	0.671	0.645	-1.67%	-0.02%	L. Andreev, “The Life of Vasily Fiveysky”
		Maximum complexity	0.742	0.672	8.69%	4.24%	D. Furmanov, “Chapayev”
		Mean entropy and complexity	0.683	0.651	0.02%	0.92%	A. Chekhov, “The Duel”
6	2	Minimum entropy	0.689	0.646	-7.99%	0.17%	Z. Gippius, “The Moon”
		Mean entropy	0.750	0.657	0.09%	1.98%	S. Bobrov, “Taurica, or My Summer Day at the Taurian Chersonesus”
		Maximum entropy	0.839	0.567	12.03%	-11.99%	I. Goncharov, “The Frigate Pallada”
		Minimum complexity	0.823	0.552	9.86%	-14.34%	L. Tolstoy, “War and Peace”
		Mean complexity	0.785	0.645	4.71%	0.05%	M. Gorky, “Childhfood”
		Maximum complexity	0.738	0.659	-1.50%	2.24%	A. Platonov, “The Juvenile Sea”
		Mean entropy and complexity	0.750	0.657	0.09%	1.98%	S. Bobrov, “Taurica, or My Summer Day at the Taurian Chersonesus”

S3 Table. Minimum, average, and maximum entropies and complexities for texts of the English language

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
3	5	Minimum entropy	0.707	0.471	-7.46%	4.20%	C. Darwin, “The Different Forms of Flowers on Plants of the Same Species”
			0.764	0.443	0.00%	-2.18%	W. Collins, “Little Novels”
		Mean entropy	0.764	0.451	0.00%	-0.27%	J. London, “The Night-Born and Other Stories”
			0.765	0.451	0.00%	-0.30%	E. Nesbit, “The Phoenix and the Carpet”
		Maximum entropy	0.802	0.398	4.93%	-11.98%	Lord Byron, “Don Juan”
		Minimum complexity	0.802	0.398	4.93%	-11.98%	Lord Byron, “Don Juan”
		Mean complexity	0.746	0.452	-2.45%	-0.02%	M. Davenport, “The Valley of Decision”
		Maximum complexity	0.718	0.503	-6.09%	11.11%	G. K. Chesterton, “The Barbarism of Berlin”
		Mean entropy and complexity	0.765	0.452	0.02%	-0.05%	W. Collins, “The Legacy of Cain”
3	6	Minimum entropy	0.659	0.531	-6.95%	-2.54%	C. Darwin, “The Different Forms of Flowers on Plants of the Same Species”
		Mean entropy	0.708	0.536	0.00%	-1.64%	N. Hawthorne, “The Scarlet Letter”
		Maximum entropy	0.746	0.507	5.31%	-7.07%	Lord Byron, “Don Juan”
		Minimum complexity	0.746	0.507	5.33%	-7.09%	Lord Byron, “Don Juan”
		Mean complexity	0.714	0.545	0.82%	-0.01%	S. White, “Blazed Trail Stories”
			0.715	0.545	0.88%	0.01%	H. MacGrath, “Parrot & Co.”
		Maximum complexity	0.694	0.580	-2.02%	6.30%	H. James, “The Marriages”
		Mean entropy and complexity	0.708	0.545	-0.05%	-0.03%	G. Keith , “The Man Who Was Thursday”
3	7	Minimum entropy	0.608	0.572	-7.72%	-2.83%	H.Wadsworth, “Evangeline”
		Mean entropy	0.658	0.587	0.00%	-0.28%	R. Barbour, “The Half-Back”
		Maximum entropy	0.709	0.585	7.68%	-0.75%	E. Spenser, “The Faerie Queene”
		Minimum complexity	0.640	0.559	-2.83%	-5.06%	A.C. Doyle, “The Great Boer War”

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
4	3	Mean complexity	0.626	0.589	-4.91%	0.00%	P. White, “Life in the War Zone”
			0.672	0.589	2.14%	0.00%	H. James, “The Golden Bowl”
			0.679	0.589	3.17%	0.00%	R. Connor, “The Patrol of the Sun Dance Trail”
		Maximum complexity	0.673	0.618	2.24%	4.86%	W. Shakespeare, “Midsummer Night's Dream”
		Mean entropy and complexity	0.658	0.589	-0.01%	-0.06%	E. Wallace, “Bones”
		Minimum entropy	0.754	0.506	-8.96%	18.11%	H.Wadsworth, “Evangeline”
		Mean entropy	0.829	0.414	0.00%	-3.36%	J. Altsheler, “The Quest of the Four”
			0.829	0.417	0.00%	-2.66%	A.B. Reeve, “The War Terror”
			0.829	0.454	0.00%	6.08%	J.M. Barrie, “Dear Brutus”
		Maximum entropy	0.858	0.381	3.59%	-10.88%	H.R. Haggard, “Dawn”
		Minimum complexity	0.853	0.362	2.99%	-15.54%	Lord Byron, “Don Juan”
		Mean complexity	0.820	0.428	-1.01%	0.00%	S. White, “African Camp Fires”
		Maximum complexity	0.754	0.506	-8.96%	18.11%	H.Wadsworth, “Evangeline”
		Mean entropy and complexity	0.828	0.428	-0.13%	0.01%	E. Wallace, “The Daffodil Mystery”
4	4	Minimum entropy	0.660	0.613	-9.98%	-1.45%	H.Wadsworth, “Evangeline”
		Mean entropy	0.733	0.628	0.01%	0.85%	A. Johnston, “Joel: A Boy of Galilee”
			0.733	0.617	0.01%	-0.82%	A. Reeve, “The Film Mystery”
		Maximum entropy	0.780	0.583	6.47%	-6.30%	Lord Byron, “Don Juan”
		Minimum complexity	0.771	0.581	5.19%	-6.67%	A. Trollope, “The Way We Live Now”
		Mean complexity	0.674	0.622	-8.01%	0.01%	N. Hawthorne, “Biographical Sketches”
		Maximum complexity	0.726	0.643	-0.97%	3.40%	W. Shakespeare, “Twelfth Night”
		Mean entropy and complexity	0.733	0.622	0.06%	0.03%	E. Wallace, “The Daffodil Mystery”
5	3	Minimum entropy	0.624	0.610	-12.04%	-7.18%	G. K. Chesterton, “The Barbarism of Berlin”
		Mean entropy	0.710	0.659	0.00%	0.20%	S. Anderson, “Marching Men”
		Maximum entropy	0.784	0.653	10.47%	-0.71%	A. Trollope, “Can You Forgive Her?”

n	d	Category	Entropy	Complexity	Mean deviation (Entropy) %	Mean deviation (Complexity) %	Literature masterpiece
		Minimum complexity	0.624	0.610	-12.04%	-7.18%	G. K. Chesterton, “The Barbarism of Berlin”
			0.698	0.657	-1.65%	0.00%	J. London, “The Strength of the Strong”
		Mean complexity	0.692	0.657	-2.52%	0.00%	E.P. Roe, “Found Yet Lost”
			0.693	0.657	-2.26%	0.00%	R. Tagore, “Mashi and Other Stories”
			0.761	0.657	7.27%	0.00%	W. Collins, “Man and Wife”
		Maximum complexity	0.738	0.673	4.04%	2.37%	H.R. Haggard, “Dawn”
		Mean entropy and complexity	0.710	0.659	0.00%	0.20%	S. Anderson, “Marching Men”
		Minimum entropy	0.672	0.630	-12.70%	-1.05%	H.Wadsworth, “Evangeline”
		Mean entropy	0.769	0.652	0.00%	2.49%	H.R. Garis, “The Curlytops on Star Island”
			0.769	0.652	0.00%	2.52%	H.R. Garis, “The Curlytops Snowed In”
			0.769	0.645	0.00%	1.38%	W. Scott, “The Black Dwarf”
		Maximum entropy	0.853	0.539	10.84%	-15.23%	A. Trollope, “Can You Forgive Her?”
6	2	Minimum complexity	0.852	0.535	10.73%	-15.88%	A. Trollope, “The Way We Live Now”
		Mean complexity	0.777	0.636	1.06%	0.00%	D.H. Lawrence, “Sea and Sardinia”
			0.794	0.636	3.20%	0.00%	A. Christie, “The Secret Adversary”
			0.786	0.636	2.15%	0.00%	T. Carlyle, “Latter Day Pamphlets”
		Maximum complexity	0.738	0.660	-4.11%	3.67%	W.S. Maugham, “Caesar's Wife”
		Mean entropy and complexity	0.769	0.639	-0.02%	0.43%	J. London, “Jerry of the Islands”

Appendix D. Entropy and complexity values for literature masterpieces and their translations.

S4 Table. Normalised entropies and complexities for “Oliver Twist” by Ch. Dickens and its translations

A translation	n	d	Entropy	Complexity	Entropy deviation from the original (%)	Complexity deviation from the original (%)
The original	3	5	0.775	0.431	0%	0%
	3	6	0.723	0.523	0%	0%

A translation	n	d	Entropy	Complexity	Entropy deviation from the original (%)	Complexity deviation from the original (%)
	3	7	0.679	0.579	0%	0%
	4	3	0.836	0.403	0%	0%
	4	4	0.761	0.600	0%	0%
	5	3	0.754	0.664	0%	0%
	6	2	0.813	0.600	0%	0%
V. Lukianskaya	3	5	0.753	0.456	-2.82%	5.82%
	3	6	0.692	0.536	-4.32%	2.52%
	3	7	0.645	0.580	-4.93%	0.18%
	4	3	0.818	0.442	-2.20%	9.65%
	4	4	0.712	0.620	-6.34%	3.37%
	5	3	0.702	0.658	-6.89%	-0.82%
	6	2	0.774	0.649	-4.86%	8.26%
E. Lann and A. Krivtsova	3	5	0.770	0.437	-0.62%	1.48%
	3	6	0.716	0.516	-0.89%	-1.23%
	3	7	0.675	0.575	-0.60%	-0.73%
	4	3	0.833	0.412	-0.38%	2.10%
	4	4	0.741	0.604	-2.62%	0.66%
	5	3	0.741	0.663	-1.71%	-0.08%
	6	2	0.813	0.613	-0.10%	2.24%
Facebook	3	5	0.751	0.473	-3.05%	9.73%
	3	6	0.688	0.555	-4.85%	6.10%
	3	7	0.637	0.588	-6.10%	1.54%
	4	3	0.809	0.469	-3.31%	16.43%
	4	4	0.693	0.626	-8.90%	4.37%
	5	3	0.669	0.642	-11.29%	-3.30%
	6	2	0.736	0.656	-9.49%	9.44%

S5 Table. Normalised entropies and complexities for "Crime and Punishment" by F. Dostoyevsky and its translations

A translation	n	d	Entropy	Complexity	Entropy deviation from the original (%)	Complexity deviation from the original (%)
The original	3	5	0.756	0.441	0%	0%
	3	6	0.700	0.513	0%	0%
	3	7	0.662	0.566	0%	0%
	4	3	0.826	0.414	0%	0%
	4	4	0.732	0.595	0%	0%
	5	3	0.741	0.657	0%	0%
	6	2	0.818	0.601	0%	0%
K. Garnett	3	5	0.778	0.432	3.04%	-2.15%
	3	6	0.726	0.524	3.64%	2.09%
	3	7	0.681	0.579	2.88%	2.39%
	4	3	0.850	0.388	2.91%	-6.31%
	4	4	0.768	0.595	4.91%	0.01%
	5	3	0.769	0.663	3.75%	0.97%
	6	2	0.833	0.579	1.77%	-29.26%
R. Pevear and L. Volkhonskaya	3	5	0.783	0.425	3.65%	-3.72%
	3	6	0.732	0.520	4.48%	1.36%
	3	7	0.688	0.579	3.93%	2.33%
	4	3	0.853	0.378	3.32%	-8.57%
	4	4	0.773	0.591	5.61%	-0.74%
	5	3	0.773	0.661	4.37%	0.66%
	6	2	0.837	0.572	2.24%	-4.87%
Facebook	3	5	0.756	0.463	0.07%	5.01%
	3	6	0.698	0.554	-0.38%	8.08%
	3	7	0.645	0.590	-2.52%	4.32%
	4	3	0.821	0.447	-0.55%	7.90%

A translation	n	d	Entropy	Complexity	Entropy deviation from the original (%)	Complexity deviation from the original (%)
	4	4	0.720	0.632	-1.65%	6.22%
	5	3	0.693	0.656	-6.49%	-0.11%
	6	2	0.753	0.654	-7.94%	8.80%
	3	5	0.731	0.443	-3.25%	0.46%
	3	6	0.681	0.517	-2.79%	0.76%
	3	7	0.649	0.561	-1.89%	-0.87%
Word-by-word	4	3	0.766	0.435	-7.21%	5.15%
	4	4	0.686	0.575	-6.30%	-3.44%
	5	3	0.683	0.622	-7.77%	-5.28%
	6	2	0.737	0.601	-9.90%	0.04%

S6 Table. Normalised entropies and complexities for “Anna Karenina” by L. Tolstoy and its translations

A translation	n	d	Entropy	Complexity	Entropy deviation from the original (%)	Complexity deviation from the original (%)
	3	5	0.761	0.430	0%	0%
	3	6	0.707	0.498	0%	0%
	3	7	0.669	0.553	0%	0%
The original	4	3	0.824	0.411	0%	0%
	4	4	0.737	0.581	0%	0%
	5	3	0.750	0.646	0%	0%
	6	2	0.827	0.578	0%	0%
	3	5	0.772	0.434	1.49%	0.99%
	3	6	0.719	0.518	1.78%	3.99%
K. Garnett	3	7	0.676	0.568	1.10%	2.80%
	4	3	0.845	0.389	2.56%	-5.35%
	4	4	0.769	0.581	4.42%	0.01%

	5	3	0.779	0.649	3.76%	0.42%
	6	2	0.840	0.547	1.61%	-5.26%
R. Pevear	3	5	0.776	0.433	1.91%	0.56%
	3	6	0.724	0.518	2.47%	4.01%
	3	7	0.681	0.570	1.89%	3.21%
	4	3	0.848	0.385	2.82%	-6.42%
	4	4	0.772	0.580	4.78%	-0.06%
	5	3	0.779	0.650	3.85%	0.56%
	6	2	0.841	0.548	1.72%	-5.06%
Facebook	3	5	0.759	0.477	-0.25%	10.96%
	3	6	0.697	0.571	-1.40%	14.74%
	3	7	0.639	0.597	-4.37%	7.93%
	4	3	0.821	0.464	-0.40%	13.00%
	4	4	0.701	0.637	-4.89%	9.70%
	5	3	0.658	0.637	-12.25%	-1.51%
	6	2	0.717	0.655	-13.26%	13.43%
Word-by-word	3	5	0.736	0.438	-3.29%	1.90%
	3	6	0.688	0.509	-2.69%	2.35%
	3	7	0.658	0.554	-1.56%	0.29%
	4	3	0.773	0.428	-6.18%	4.20%
	4	4	0.694	0.566	-5.83%	-2.59%
	5	3	0.697	0.619	-7.11%	-4.29%
	6	2	0.750	0.585	-9.33%	1.32%