Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2023"

June 14–16, 2023

Autocorrelations Decay in Texts and Applicability Limits of Language Models

Nikolay Mikhaylovskiy
Higher IT School, Tomsk State University, Tomsk, Russia, 634050
NTR Labs, Moscow, Russia, 129594
nickm@ntr.ai

Ilya Churilov
NTR Labs, Moscow, Russia, 129594
ichurilov@ntr.ai

Abstract

We show that the laws of autocorrelations decay in texts are closely related to applicability limits of language models. Using distributional semantics we empirically demonstrate that autocorrelations of words in texts decay according to a power law. We show that distributional semantics provides coherent autocorrelations decay exponents for texts translated to multiple languages. The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. We conclude that language models exhibiting Markovian behavior, including large autoregressive language models, may have limitations when applied to long texts, whether analysis or generation.

Keywords: autocorrelations decay laws, language models
DOI: 10.28995/2075-7182-2023-22-350-360

Убывание автокорреляций в текстах и границы применимости языковых моделей

Николай Михайловский
Высшая ИТ-Школа Томского Государственного Университета, Томск, Россия, 634050
ООО «НТР», Москва, Россия, 129594
nickm@ntr.ai

Илья Чурилов
ООО «НТР», Москва, Россия, 129594
ichurilov@ntr.ai

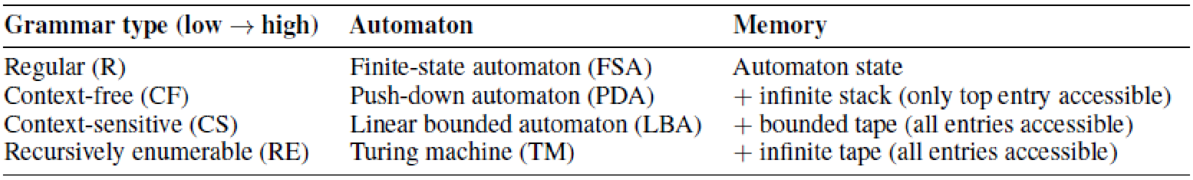
Аннотация

Показано, что законы затухания автокорреляций в текстах тесно связаны с пределами применимости языковых моделей. С использованием дистрибуционной семантики продемонстрировано, что автокорреляции слов в литературных текстах затухают по степенному закону. Показано, что дистрибуционная семантика обеспечивает когерентные показатели затухания автокорреляций для текстов, переведенных на несколько языков. Затухание автокорреляций в сгенерированных текстах количественно и часто качественно отличается от художественных текстов. Таким образом, языковые модели, демонстрирующие марковское поведение, включая большие авторегрессионные языковые модели, могут иметь ограниченную применимость к длинным текстам, будь то анализ или генерация.

Ключевые слова: большие языковые модели, законы убывания автокорреляции

1 Introduction

In this work, we endeavor into outlining statistically the limits of applicability of popular contemporary language models. To avoid any terminological doubt, when we write "models of the language", we refer to any models that explain some linguistic phenomena, while "language models" refer to probabilistic



Grammar type (low → high)	Automaton	Memory
Regular (R)	Finite-state automaton (FSA)	Automaton state
Context-free (CF)	Push-down automaton (PDA)	+ infinite stack (only top entry accessible)
Context-sensitive (CS)	Linear bounded automaton (LBA)	+ bounded tape (all entries accessible)
Recursively enumerable (RE)	Turing machine (TM)	+ infinite tape (all entries accessible)

Table 1: Chomsky hierarchy of formal grammars (from [10])

language models as defined in Subsection 2.3 Probabilistic Language Models. While not long ago probabilistic language models were just models that assign probabilities to sequences of words [4], now they are the cornerstone of any task in computational linguistics through few-shot learning [6], prompt engineering [38] or fine-tuning [13]. On the other hand, current language models fail to catch long-range dependencies in the text consistently. For example, text generation with maximum likelihood target leads to rapid text degeneration, and consistent text generation requires probabilistic sampling and other tricks [22]. Large language models such as GPT-3 [6] push the boundary of “short text” rather far (specifically, to 2048 tokens), but do not remove the problem.

Our contributions in this work are the following:

- We explain how the laws of autocorrelations decay in texts are related to applicability of language models to long texts;
- We pioneer the use of pretrained word vectors for autocorrelation computations that allows us to study a widest range of autocorrelation distances;
- We show that the autocorrelations in literary texts decay according to power laws for all these distances;
- We show that distributional semantics typically provides coherent autocorrelations decay exponents for texts translated to multiple languages, unlike earlier flawed approaches;
- We show that the behavior of autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts.

2 Models of the Language

In this section, we briefly introduce models of the language that are important for the further considerations.

2.1 Formal Grammars

Formal grammars describe how to form strings from a language's alphabet that are valid according to the language's syntax. They were introduced by Chomsky in 1950s [7][8]. A formal grammar consists of a finite set of production rules in the form

$$left - hand\ side \ \rightarrow \ right - hand\ side, \tag{1}$$

where each side consists of a finite sequence of the following symbols:

- a finite set of nonterminal symbols (indicating that some production rule can yet be applied)
- a finite set of terminal symbols (indicating that no production rule can be applied)
- a start symbol (a distinguished nonterminal symbol)

Chomsky grammars constitute a hierarchy, see Table 1. While the original hierarchy implies strict inclusion of lower class grammars to higher ones, now there are several types of grammars known to fall between or partially overlap with the original classes (see, for example, [10]).

2.2 Distributional Semantics and Models

Distributional hypothesis assumes that linguistic items with similar distributions have similar meanings or function and was likely first introduced by Harris [20] in 1954 and was popularized in the form "a word is characterized by the company it keeps" by Firth [17]. The basic idea is to collect distributional information in, say, high-dimensional vectors, and then to define similarity in terms of some metric, say Euclidean distance or the angle between the vectors.

Early distributional approaches from 60s relied on hand-crafted features of the words [35], while more recent – on statistics of varied sorts. The first generation of statistical distributional semantics approaches included Latent Semantic Analysis (LSA) [11][12], Hyperspace Analogue to Language (HAL) [24][25], and many others, see [15] for a review. The second generation primarily consists of word2vec [31][32] and GloVe [37] models, the first, implicitly, and the second, explicitly adding the word analogy task into the training objective, so that similar relationships between words should be described by similar difference vectors between embeddings. The third generation of statistical distributional semantics models was started by emergence of BERT contextual word embeddings [13]. BERT have combined the word and its current context into a single vector embedding and used Masked Language Modelling training objective. A lot of recent work sprouted from BERT.

2.3 Probabilistic Language Models

Probabilistic language models consider sequences

$$t_{1:m} = \{t_1, t_2, \dots, t_m\} \tag{2}$$

of tokens from the lexicon $\mathcal{L}$. An autoregressive language model estimates the probability of such a sequence

$$P(t_{1:m}) = P(t_1)P(t_2|t_1)P(t_3|t_{1:2}) \dots P(t_m|t_{1:m-1}) = \prod_{k=1}^{m} P(t_k|t_{1:k-1}) \tag{3}$$

using the chain rule. Most models introduce the Markov [30] assumption that the probability of a token depends on the previous $n-1$ tokens only, thus approximating (3) with a truncated version

$$P(t_{1:m}) \approx \prod_{k=1}^{m} P(t_k|t_{k-n+1:k-1}) \tag{4}$$

While the language models based on recurrent [33], and specifically, LSTM [41] neural networks do not introduce the Markov assumption explicitly, we will shortly see that in practice they do exhibit Markovian behavior. On the other hand, it is long known that Markov models describe stochastic regular grammars [42].

3 Why Autocorrelations Decay Laws Matter?

In this section we explain why autocorrelation decay laws matter a lot to computational linguistics' near-future.

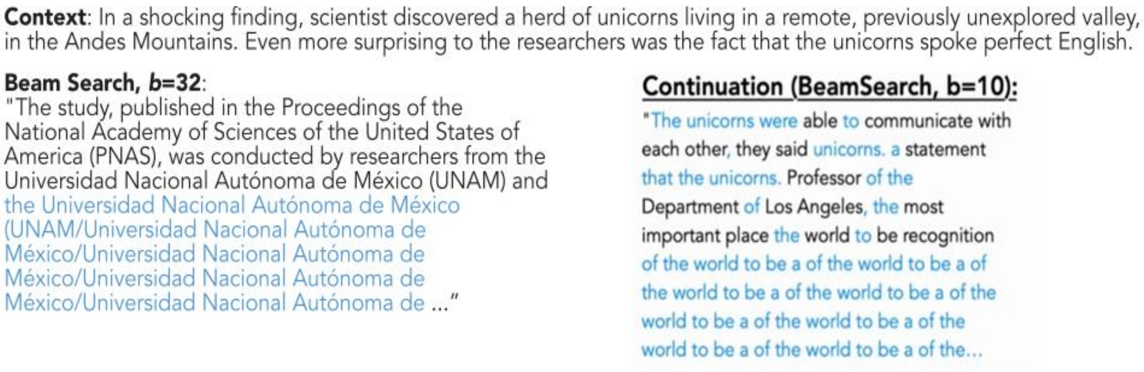
3.1 Computing Autocorrelations Using Distributional Semantics

Suppose we have a sequence of N vectors $V_i \in R^d, i \in [1, N]$. Autocorrelation function $C(\tau)$ is the average similarity between the vectors as a function of the lag $\tau = i - j$ between them. The simplest metric of vector similarity is the cosine distance $d(V_i, V_j) = \cos\angle(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|}$, where $\cdot$ is a dot product between two vectors and $\| \quad \|$ is an Euclidean norm of a vector. Thus,

$$C(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \frac{V_i \cdot V_{i+\tau}}{\|V_i\| \|V_{i+\tau}\|}. \tag{5}$$

$C(\tau)$ ranges from -1 for perfectly anticorrelated sequence (for $\tau = 1$ and $d = 1$ that would be $1, -1, 1, -1$ etc.) to 1 for a perfectly correlated one (for $\tau = 1$ and $d = 1$ that would be $1, 1, 1, 1$ etc.).

A distributional semantic assigns a vector to each word or context in a text. Thus, a text is transformed into a sequence of vectors, and we can calculate an autocorrelation function for the text.



Context: In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Beam Search, *b*=32:
"The study, published in the Proceedings of the National Academy of Sciences of the United States of America (PNAS), was conducted by researchers from the Universidad Nacional Autónoma de México (UNAM) and the Universidad Nacional Autónoma de México (UNAM/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de México/Universidad Nacional Autónoma de ..."

Continuation (BeamSearch, b=10):
"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the...

Figure 1: Beam search produces degenerate text (from [22])

Level	Architecture	Description
R-	Transformer	The encoder with stacked multi-head attention layers and dense layers.
R	RNN	A classical RNN with ReLU activations.
R+	LSTM	A classical LSTM.
DCF+	Stack-RNN	An RNN with an external stack, with PUSH, POP, and NO-OP actions.
NDCF	NDStack-RNN	An RNN with a nondeterministic stack, simulated using dynamic programming
CS	Tape-RNN	An RNN with a finite tape, as in a Turing machine (similar to Baby-NTM).

Table 2: Alignment of neural network architectures with Chomsky hierarchy (from [10])

3.2 Transformer Language Models Exhibit Markovian Behavior

In this paper, by Markovian behavior, we mean that large language models actually use only a limited context, often significantly less than the maximum context possible. Thus they implicitly or explicitly use the Markov assumption. Two separate phenomena classes that prove that transformer language models exhibit Markovian behavior are known, and in Section 5.5 we introduce the third one.

One such phenomenon is the rapid text degeneration when a transformer language model is used to generate text with maximum likelihood target [21][28]. Maximization-based decoding methods such as beam search lead to output text that is bland, incoherent, or gets stuck in repetitive loops [22] that are extremely reminiscent of positively recurrent Markov chains (see Figure 1).

The other phenomenon is studied in detail in [10]. The authors have found that the networks roughly match the computational models associated with the Chomsky hierarchy: RNNs can solve tasks up to the regular level, Stack-RNNs up to the DCF level, and Tape-RNNs up to the CS level. Finally, they observed that Transformers and LSTMs are less aligned with the Chomsky hierarchy: Transformers fail on regular tasks, while LSTMs can solve tasks more difficult than regular. The results of [10] are summarized in Table 2. As transformers can at most generalize to regular languages and Markov models describe stochastic regular grammars [42], we can safely say that transformers exhibit behavior no richer than regular.

3.3 Markovian Implies Exponential Correlations Decay, Probabilistic Context-Free Grammars Can Generate Power Laws

Assume that the sequence (2) is an output of a random source that takes values in $\mathcal{L}$. If the source is Markovian, it can be shown [23] that the autocorrelations (or, equivalently, mutual information between chunks of the text) decay exponentially. Namely, the following theorem holds:

Theorem 1 ([23]). Let M be a Markov matrix that generates a Markov process. If M is irreducible and aperiodic, then the asymptotic behavior of the mutual information $I(t_1, t_2)$ is exponential decay toward zero for $|t_2 - t_1| \gg 1$ with decay timescale $\log\frac{1}{|\lambda_2|}$, where λ_2 is the second largest eigenvalue of M. If M is reducible or periodic, I can instead decay to a constant; no Markov process whatsoever can produce power law decay.

One the other hand, the following theorem holds:

Theorem 3 ([23]). There exist a probabilistic context-free grammar such that the mutual information $I(A, B)$ between two symbols A and B in the terminal strings of the language decay like d^{-k}, where d is the number of symbols between A and B.

3.4 If the Natural Language Exhibits Power Law Correlations Decay, We Can Do Better Than Autoregressive Language Models

Summarizing the above, if texts in the natural languages exhibit exponential autocorrelations decay, autoregressive language models are good to analyze or generate texts of any length. On the other hand, if texts in the natural languages exhibit power law autocorrelations decay, building language models that exhibit at least hierarchical, context-free-grammar-ish, slow-correlation-decay behavior may be beneficial for a variety of downstream tasks. This may be not enough to model long texts successfully because natural languages cannot be completely described by context-free grammars (see, for example, [40]), but may be a meaningful step.

4 Studying Autocorrelations Decay Laws in Texts

4.1 Prior Research

Schenkel, Zhang, and Zhang [39] were likely the first to empirically find the power law autocorrelations decay in texts using a random walk model with an arbitrary mapping of characters to fixed length, 5 bit sequences. They studied 10 texts in English. The obvious drawback of their approach is dependency on encoding. Amit et al. [3] explored this problem in various translations of the Bible and have shown that the power law exponent depends on both the language and the codification. Testing multiple random mappings would provide a more reliable estimate of power law exponents, but such a research is a matter of future. Random walk models have later been used to find the power law in text by several researchers, including Ebeling and Neiman [14], Kokol et al. [26] (who, by the way, in our opinion have not found power-law autocorrelations in literary writing on distances studied, but found power-law autocorrelations in computer programs, in a perfect agreement with the fact that computer programs are described by context-free grammars), Pavlov et al. [36], who find multifractal structures in the text, and Manin [29], who attribute long-range correlations to slow variations in lexical composition within the text.

Alvarez-Lacalle et al. [2] used a version of first-generation distributional semantic model to study autocorrelations in 12 literary texts in English to find power law autocorrelations decay. Altmann, Cristadoro, and Degli [1] analyze 41 binary functions on words separately on ten English versions of international novels. They separate the effects of burstiness and long-range correlations in the power spectrum and find a power law correlations decay. Lin and Tegmark [23] in a short empirical part of their study use three text corpora: 100 MB from Wikipedia, the first 114 MB of a French corpus and 27 MB of English articles from slate.com. They observe the power law decay of mutual information, but note that the large portion of the long-range mutual information appears to be dominated by poems in the French sample and by the html-like syntax in the Wikipedia sample. They have also shown similar power decay laws for autocorrelations in natural music and exponential laws in generated music, the result reproduced by different means by Yamshchikov and Tikhonov [43]. Corral et al. [10] study intervals between consecutive appearances of specific words in literary texts in 4 languages, including Finnish (a rare study of highly agglutinative language) to find that most words have a universal dimensionless probability density function described by gamma distribution. Gillet and Ausloos [18] and Montemurro and Pury [34] study sequences built from word frequencies and word lengths to find the power law autocorrelations decay.

4.2 Research Questions

Given the prior art, many research question remain unanswered. The ones we address in this work are:

Q1. How accurately can we say that autocorrelations in texts decay according to a power law?

Q2. Can we reject the hypothesis of exponential decay of correlations?

Q3. Does the law of decay depend on the language of the text?

Q4. Over what range of distances does the decay in autocorrelations follow a power law?

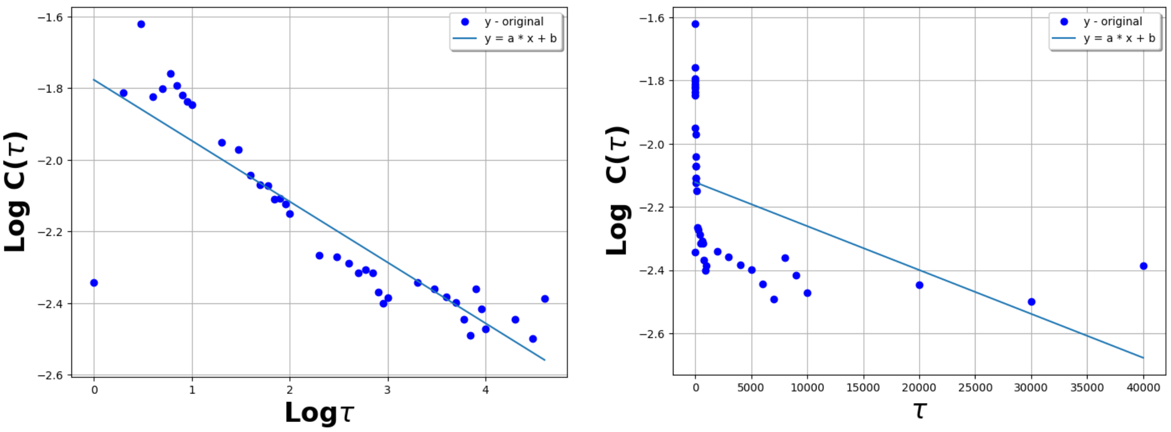


Figure 2: Autocorrelations in Don Quixote (English) computed using GloVe, a = 1, d = 300, $\tau \in [1, 40000]$ Left: log-log coordinates. Right: log-linear coordinates.

Q5. Are autocorrelations in LM-generated texts any different from literary texts?

4.3 Methods

In this work we use two distributional semantic models to estimate autocorrelations in long texts. One is a bag-of-words (BOW) embedding model of Alvarez-Lacalle et al. [2]. The other distributional semantic model we use is GloVe [37]. We use pretrained multilingual GloVe vector embeddings from [16]. We filter out both frequent and rare words filtering similarly to [2] when using BOW.

BOW assigns a vector of dimension d to each word first, and then averages these vectors over a window of the size a. This averaged vector is then assigned to a word in the center of averaging window. The exact procedure for BOW is described in detail in [2]. GloVe naturally maps each word to a vector; we then center the vector system by subtracting the average of vectors over the whole text, and, similarly, average over a window of the size a when we need direct comparison to BOW. After that in both cases we can compute the autocorrelation function following Section 3.1.

5 Experiments

5.1 The Dataset

For our studies we have collected a dataset of long literary and philosophical works in English, Spanish, French, German and Russian[i] each: Critique of Pure Reason, Don Quixote de la Mancha, Moby-Dick or, The Whale, The Adventures of Tom Sawyer, The Iliad, The Republic and War and Peace. The only translation absent is Moby-Dick in German, which happened to be substantially abridged. The texts have been obtained from Project Gutenberg, Wikisource, Royallib and lib.ru and preprocessed so as to fit our research purposes:

- copyright texts were removed from the files;
- author and translator notes were removed;
- table of contents and any indices were removed, except for the table of contents from Don Quixote;
- any links to illustrations have been removed;
- in the Russian version of War and Peace any non-Russian text have been replaced with Russian translations;
- etymology section was removed from Moby-Dick or, The Whale, where encountered, as some languages missed it.

5.2 Choosing Between Hypotheses of Power Law and Exponential Decay of Correlations

To address **Q1.** "How accurately can we say that autocorrelations in texts decay according to a power law?" and **Q2.** "Can we reject the hypothesis of exponential decay of correlations?" for each text, we

	Power Law					Exponential Law				
	BOW en	fr	es	ru	en	BOW en	fr	es	ru	en
The Adventures of Tom Sawyer	**0,16**	**0,11**	**0,16**	**0,14**	**0,21**	0,52	0,32	0,33	0,33	0,55
The Republic	**0,21**	**0,15**	**0,09**	**0,10**	**0,13**	0,58	0,28	0,25	0,31	0,38
Don Quixote	**0,20**	**0,11**	**0,12**	**0,09**	**0,20**	0,66	0,24	0,22	0,23	0,44
War and Peace	**0,20**	**0,13**	**0,11**	**0,08**	**0,09**	0,54	0,24	0,24	0,28	0,42
Critique of Pure Reason	**0,09**	**0,07**	**0,15**	**0,10**	**0,14**	0,27	0,17	0,20	0,21	0,25
The Iliad	**0,24**	2,37	**0,16**	**0,10**	**0,19**	0,63	**2,33**	0,17	0,19	0,54
Moby-Dick or, The Whale	**0,14**	**0,12**	**0,11**	**0,09**	**0,15**	0,40	0,22	0,22	0,22	0,47

Table 3: Goodness of fit of autocorrelation by power and exponential laws in terms of MAPE. BOW: a=200, d=100, $\tau \in [250, 4200]$ Glove: a = 1, d = 300, $\tau \in [\varepsilon, 40000]$

	BOW			GloVe		
	α	β	MAPE	α	β	MAPE
en	-0.7718	0.9545	0.1054	-0.7246	1.1582	0.1044
fr	-0.8836	1.1407	0.2154	-0.7749	1.1051	0.2150
es	-0.7601	0.9332	0.1057	-0.7083	0.9947	0.1271
ru	-0.7412	0.7874	0.0787	-0.6431	0.9173	0.0548
de	-0.8072	0.9542	0.1411	-0.8326	1.3478	0.1657

Table 4: Dependence of the autocorrelations power decay law in Don Quixote on the language and embedding. τ ranges from 200 to 4000 words, d=300, a = 200

have computed autocorrelations for a series of distances $\tau = n * 10^k, n \in [1, 9]$, and approximated the points produced by a straight line in both log-log and log-linear coordinates using the least squares regression. We have evaluated the goodness of fit of each model by MAPE (Mean Absolute Percentage Error). The range of τ for Glove was chosen from the first non-negative autocorrelation value ε (autocorrelations on small distances $\tau = [1, 2]$ happened to be sometimes negative).

The results for the English translation of Don Quixote are presented in the Figure 2. It can be seen that in log-log coordinates the regressed straight line approximates data well enough, unlike log-linear coordinates.

Table 3 lists the MAPE metrics of goodness of fit of autocorrelation by power and exponential laws (the smaller the better). It can be easily seen that for all the texts but one the hypothesis of exponential decay of correlations can be rejected. The peculiarity of the French translation of The Iliad is that the autocorrelation with $\tau = 1$ is very small but still positive, thus both producing significantly larger MAPE and ruining the approximation. Additional graphs are presented in the Appendix A.

5.3 Determining the Dependency of the Autocorrelations Decay Law on the Language of the Text

To study the dependency of the autocorrelations decay law on the language of the text, we have measured $C(\tau)$ for the same multilingual dataset as in Section 5.1 and fitted with power law $C(\tau) = \beta \cdot \tau^{\alpha}$. Table 4 presents results for Don Quixote. It can be easily seen that the parameters of power law, as well as the accuracy of the approximation are extremely consistent among languages for both embeddings, with standard deviation of exponent being 7% for BOW and 10% for GloVe. Moreover, the exponents for BOW and GloVe are also consistent within 15%, which we consider a very good agreement. This is in a stark contrast with the results from [3] that critically depend on the codification and language.

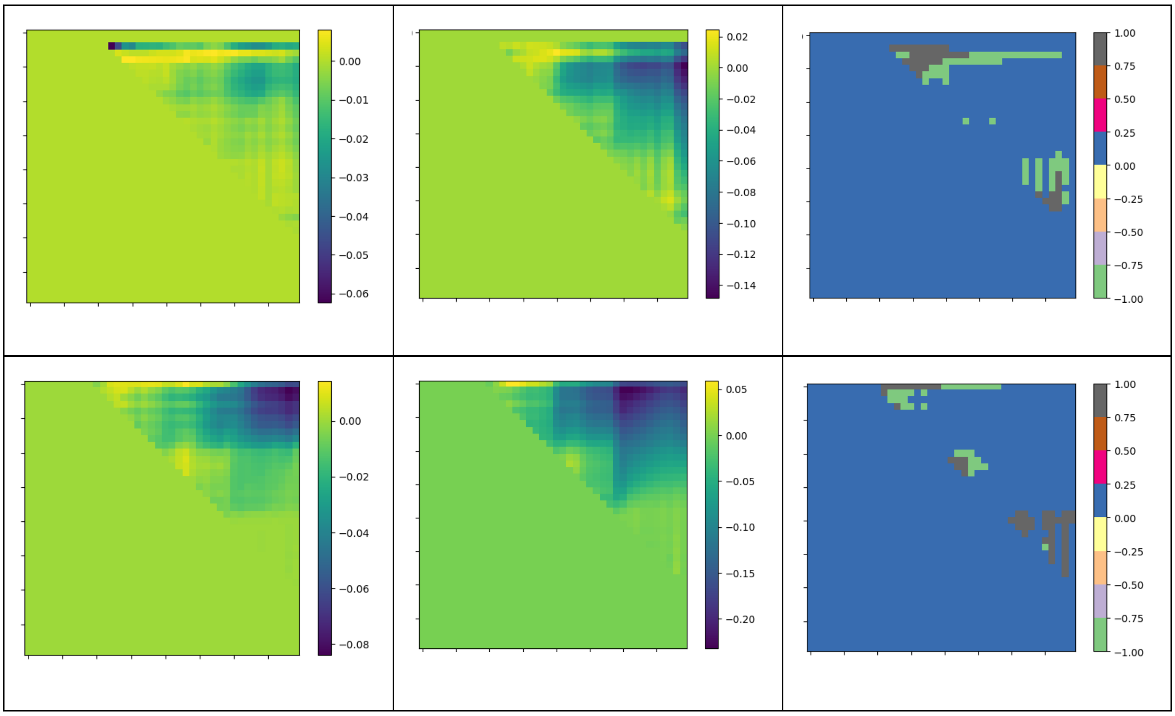


Figure 3: Autocorrelations in Critique of Pure Reason in English (top) and The Adventures of Tom Sawyer in Spanish (bottom) computed using GloVe, $a = 1, d = 300$. Vertical axis: start of τ range. Horizontal axis: end of τ range. Left: difference between power and log approximation MAPE. Middle: difference between power and exp approximation MAPE. Right: ranges where power (blue), exp (gray), and log (green) approximations are the best.

5.4 Determining the Range of Distances Where the Decay in Autocorrelations Can Be Considered Subject to a Power Law

As the BOW approach requires a sufficiently large window size a, we have studied the dependence of autocorrelations on distance ranges using GloVe embeddings with a window size $a = 1$. For each text we explored all the ranges of τ spanning at least a decimal order of magnitude, and fitted the autocorrelations with the best fitting log, power and exponential functions. We then mapped the differences between MAPE of power and other approximations, as well as the ranges where each function fits the data the best. The results for the Critique of Pure Reason in English and The Adventures of Tom Sawyer in Spanish are presented on Figure 3. Each small square on these images corresponds to a range of τ determined by its vertical (start) and horizontal (end) coordinates, for example, the full range of $\tau \in [1, 40000]$ corresponds to the top right corner. Additional graphs are presented in Appendix B.

It can be seen that for the shorter spans of τ the best approximations are sometimes logarithmic or exponential but their advantage is not significant, while for the longer ranges the best approximations are always power law. Additionally, the location of such ranges is hectic. We conclude that the cases where exponential or logarithmic approximation is better than the power law approximation represent natural short-range variability and cannot be considered a regularity.

5.5 Autocorrelations in Generated Texts

The behavior of autocorrelations is qualitatively different when the text is generated. The simplest way to generate an incoherent text is to shuffle words in a text. Figure 5 demonstrates that there is no specific autocorrelations decay law for an incoherent text.

To study autocorrelations in texts generated by large language models, we have used GPT-2 base [6] with the default generation parameters, and Structured State Space model S4 base [19] with the default generation parameters, and generated some 10K word continuous text with each model. The generated texts are listed in Appendix C and Appendix D, respectively. We then performed the same procedure as

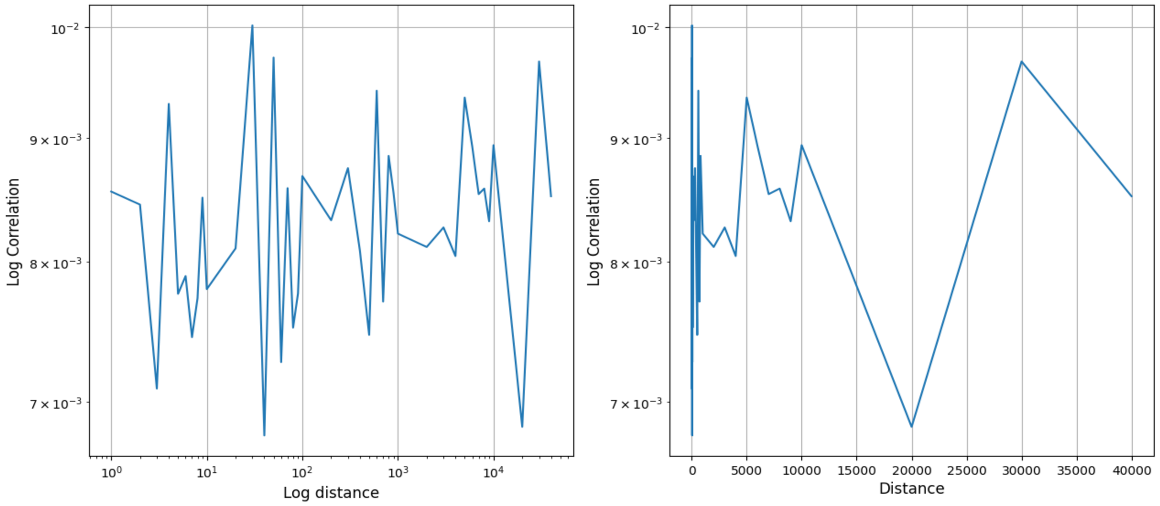


Figure 5: Autocorrelations in a randomly shuffled The Adventures of Tom Sawyer in Spanish computed using GloVe, a=1, d=300. Left: log-log, to right: log-linear coordinates

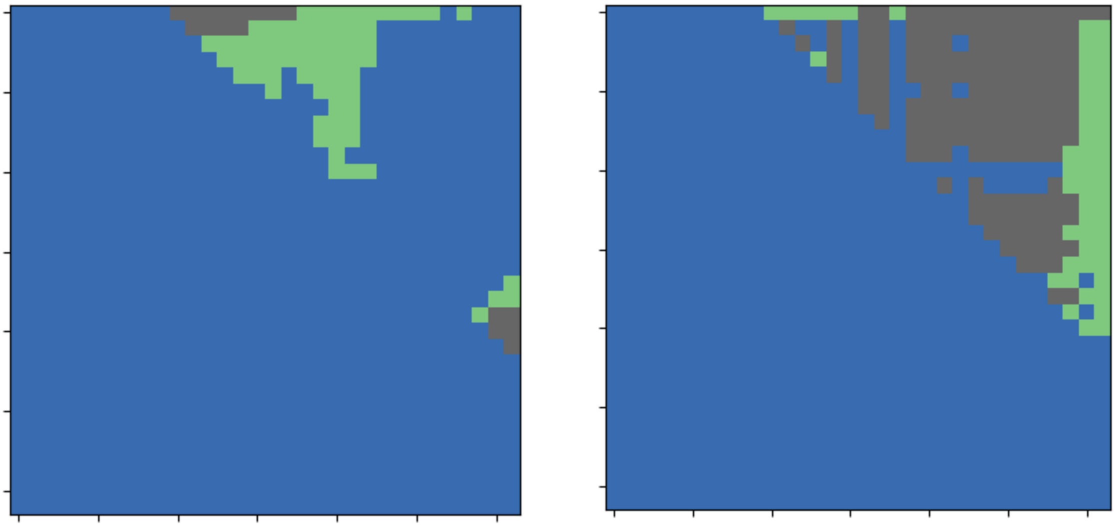


Figure 4: Autocorrelations in texts generated by GPT-2 (left) and S4 (right) models computed using GloVe, $a = 1, d = 300$, ranges where power (blue), exp (gray), and log (green) approximations are the best depicted. Vertical axis: start of τ range. Horizontal axis: end of τ range.

in Section 5.4, mapping ranges where each decay law provides the best approximation. The results are presented on Figure 4.

The autocorrelations decay in an exponential manner in the text generated by S4 model, while according to a power law on long distances and to log law – on short distances in the text generated by GPT-2. The autocorrelations in generated texts are significantly larger and decay much slower than the ones in the natural texts. In our S4 and GPT-2 generated examples, the power law coefficients are $a = -0.045, b = -0.71$ and $a = -0.027, b = -0.77$, respectively. At the same time we have not seen the coefficient a less than 0.1 for any natural text in English we have studied, and the average is closer to 0.2, indicating almost 10-fold gap between the power law decay rates in natural and generated texts. Typical values of coefficient b for natural texts are between -1.5 and -2, indicating at least 2-fold gap between natural and generated texts.

Thus we can say that the autocorrelations decay in generated texts are quantitatively and often qualitatively different from the literary texts. The conditions that influence the autocorrelations decay laws in generated texts may include sampling approach, temperature and other hyperparameters. This is a matter of future research.

6 Conclusions

We have shown empirically that autocorrelations in literary texts are decaying following the power law with the only upper limit being the length of the work itself and the hypothesis of exponential decay can be rejected for these distances. We have also shown empirically that the laws of autocorrelation decay, if measured using distributional semantics models are typically the same for the literary work translated to different languages. This contrasts previous findings that used flawed technique based on encoding-dependent random walks. Thus, we believe that distributional semantics models are a robust enough tool to measure autocorrelations in long texts.

The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. Based on the above, we can conclude that for long text processing one may need architectures different from the autoregressive ones, and many questions remain unanswered.

Acknowledgements

The authors are grateful to their colleagues at NTR Labs ML division for the discussions and support. Early versions of this work were discussed with Anton Kolonin, Dmitry Manin and Alexey Tikhonov. These discussions have improved our approach and research design, for which we are very grateful. We are also extremely grateful to Tatiana Sherstinova who discussed early versions of this work, suggested numerous improvements and provided a webinar venue at HSE to discuss this work publicly.

References

[1] Altmann E. G., Cristadoro G., Degli M. On the origin of long-range correlations in texts // PNAS. 2012. № 29 (109). C. 11582–11587.

[2] Alvarez-Lacalle E. et al. Hierarchical structures induce long-range dynamical correlations in written texts // PNAS. 2006. № 21 (103). C. 7956–7961.

[3] Amit M., Shmerler Y., Eisenberg Eli, Abraham M., Shnerb N.. Language and codification dependence of long-range correlations in texts // Fractals. 2012, №. 01 (02), C. 7-13.

[4] Bahl L.R., Jelinek F., Mercer R.L. A Maximum Likelihood Approach to Continuous Speech Recognition // IEEE Trans. Pattern Anal. Mach. Intell. 1983. Vol. PAMI-5, № 2. P. 179–190.

[5] Beltagy Iz, Peters M. E., Cohan A. Longformer: The Long-Document Transformer // arXiv:2004.05150

[6] Brown T.B. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. 2020. Vol. 2020, P. 1877–1901.

[7] Chomsky N. Three models for the description of language // IRE Transactions on Information Theory, 1956. № 2 (3): P. 113–124.

[8] Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957.

[9] Corral A. et al. Universal Complex Structures in Written Language // Vol. arXiv:0901.2924v1, Access mode: https://arxiv.org/abs/0901.2924v1

[10] Delétang G. et al. Neural Networks and the Chomsky Hierarchy // International Conference on Learning Representations, 2023

[11] Deerwester S. C. et al. Improving information retrieval using latent semantic indexing. // Proceedings of the 51st Annual Meeting of the American Society for Information Science 1988, №25, P. 36–40.

[12] Deerwester S. C. et al. Indexing by latent semantic analysis // Journal of the American Society for Information Science. №41 (6), P. 391–407.

[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding // NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. P. 4171–4186.

[14] Ebeling W., Neiman A. Long-range correlations between letters and sentences in texts // Physica A: Statistical Mechanics and its Applications, 1995, № 215 (3), P. 233-241

[15] Erk K. Vector space models of word meaning and phrase meaning: A survey // Language and Linguistics Compass, 2012, № 6(10), P. 635–653

[16] Ferreira D.C., Martins A.F.T., Almeida M.S.C. Jointly learning to embed and predict with multiple languages // 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap. 2016. Vol. 4. P. 2019–2028.

[17] Firth, J.R. A synopsis of linguistic theory 1930-1955 // Studies in Linguistic Analysis, 1957, P. 1-32. Oxford: Philological Society.

[18] Gillet J., Ausloos M. A Comparison of natural (English) and artificial (Esperanto) languages. A Multifractal method based analysis // Vol. arXiv:0801.2510, Access mode: http://arxiv.org/abs/0801.2510

[19] Gu A., Goel K., Re C. Efficiently Modeling Long Sequences with Structured State Spaces // International Conference on Learning Representations. 2021, P. 1–32.

[20] Harris, Z. Distributional structure // *Word*, 1954, №10(23), P. 146-162.

[21] Holtzman A. et al. Learning to write with cooperative discriminators // ACL 2018 - 56th Ann. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.) 2018. Vol. 1. P. 1638–1649.

[22] Holtzman A. et al. The curious case of neural text degeneration // Proceedings of the 2020 International Conference on Learning Representations. 2020. P. 2540.

[23] Lin H.W., Tegmark M. Critical behavior in physics and probabilistic formal languages // Entropy. 2017. Vol. 19, № 7. P. 1–25.

[24] Lund, K., Burgess, C., Atchley, R. A. Semantic and associative priming in a high-dimensional semantic space // Cognitive Science Proceedings (LEA), 1995, P. 660-665.

[25] Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behav. Res. Methods, Instruments, Comput. 1996. Vol. 28, № 2. P. 203–208.

[26] Kokol P. et al. Computer and natural language texts – a comparison based on long range correlations // J. Am. Soc. Inf. Sci. 1999. Vol. 50, № 14. P. 1295–1301.

[27] Kolchinsky A., Wolpert D.H. Semantic information, autonomous agency and non-equilibrium statistical physics // Interface Focus. 2018. Vol. 8, № 6

[28] Kulikov I. et al. Importance of search and evaluation strategies in neural dialogue modeling // INLG 2019 - 12th Int. Conf. Nat. Lang. Gener. Proc. Conf. 2019. P. 76–87

[29] Manin D.Y. On the nature of long-range letter correlations in texts // Vol. arXiv:0809.0103. Access mode: http://arxiv.org/abs/0809.0103. 2008. № 1. 1–14 p.

[30] Марковъ А.А. ПримЂръ статистическаго изслЂдованiя надъ текстомъ “Евгенiя ОнЂгина”, иллюстрирующiй связь испытанiй въ цЂпь // ИзвЂстiя Императорской Академiи Наукъ. VI серiя. 1913. Vol. 7, № 3. P. 153–162. In Russian. (English translation: Andrei Markov. 2006, An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Science in Context. 2006. Vol. 19, no. 4.* pages 591–600. DOI 10.1017/S0269889706001074.)

[31] Mikolov T. et al. Efficient estimation of word representations in vector space // 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR, 2013.

[32] Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality. // Proceedings of NIPS, 2013.

[33] Mikolov T. et al., Recurrent neural network based language model // Proc. of Interspeech 2010, pp. 1045–1048

[34] Montemurro M.A., Pury P.A. Long-range fractal correlations in literary corpora // Fractals. 2002. Vol. 10, № 4. P. 451–461.

[35] Osgood C., Suci G., Tannenbaum P. The measurement of meaning. — University of Illinois Press, 1957

[36] Pavlov A.N. et al. Scaling features of texts, images and time series // Phys. A Stat. Mech. its Appl. 2001. Vol. 300, № 1–2. P. 310–324.

[37] Pennington J., Socher R., and Manning C. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, P. 1532–1543.

[38] Sanh V. et al. Multitask Prompted Training Enables Zero-Shot Task Generalization // ICLR. 2022.

[39] Schenkel A., Zhang J., Zhang Y.-C. Long Range Correlation In Human Writings. // Fractals. 1993. Vol. 4, № 3. P. 229–241.

[40] Shieber S.M. Evidence against the context-freeness of natural language // Linguist. Philos. 1985. Vol. 8, № 3. P. 333–343.

[41] Sundermeyer M., Schlüter R., Ney H. LSTM neural networks for language modeling // 13th Annu. Conf. Int. Speech Commun. Assoc. 2012, INTERSPEECH 2012. 2012. Vol. 1. P. 194–197.

[42] Thompson R., Booth T., Applying Probability Measures to Abstract Languages // IEEE Transactions on Computers, 1973, vol. 22, no. 05, pp. 442-450.

[43] Yamshchikov, I.P., Tikhonov, A. Music generation with variational recurrent autoencoder supported by history // SN Appl. Sci. 2, 1937, 2020. https://doi.org/10.1007/s42452-020-03715-w

[i] The dataset is available at https://github.com/nickm197/Longtexts