Multiscale phenomena in languages and language models

Nikolay Mikhaylovskiy [1, 2][0000-0001-5660-0601]

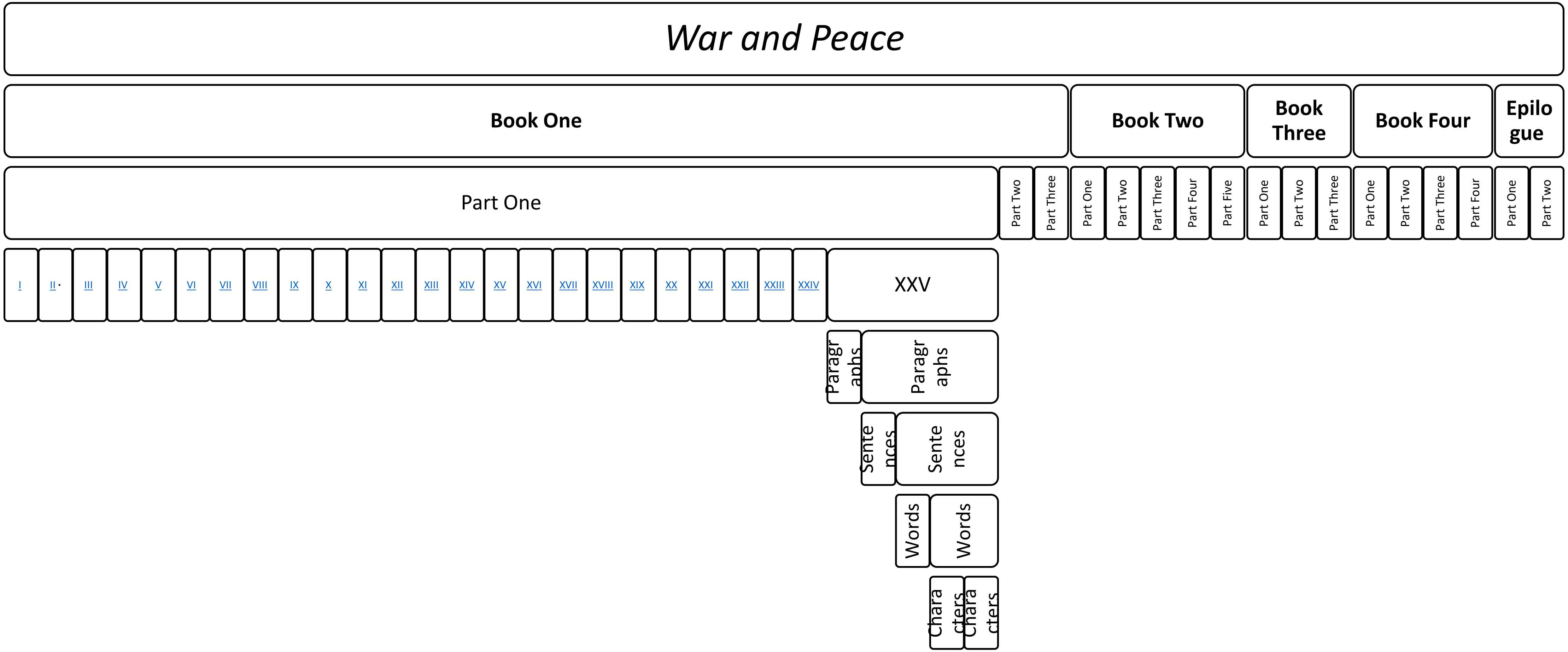
[1] Higher IT School, Tomsk State University, Tomsk, Russia, 634050

[2] NTR Labs, Moscow, Russia, 129594
nickm@ntr.ai

Multiscale structure of texts

Phenomenon under consideration

Multiscale Structure



Multiscale Structure

How human writes long text:
Core concepts ➡️ writing structure ➡️ wording and phrasing

Xiang Lisa Li. Diffusion Models for Text. Beyond autoregressive language modeling.
https://web.stanford.edu/class/cs224u/slides/lisa-224u-diffusion.pdf

Models of the language

Three established types of models of the language

- Generative formal grammars
- Distributional semantics
- Probabilistic language models, traditionally autoregressive

- Newcomer: Diffusion probabilistic language models

Generative grammars

- There is a small set of rules (grammar) that allows one to generate all the grammatical sentences of the language and not generate any ungrammatical ones.
- A formal grammar consists of a finite set of production rules in the form

$$left - hand\ side \ \rightarrow\ right - hand\ side$$

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, Noam. *Syntactic Structures*, Berlin, New York: De Gruyter Mouton, 2002.

The Chomsky Hierarchy

Grammar type (low $\rightarrow$ 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)

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg, Pedro A. Ortega. Neural Networks and the Chomsky Hierarchy, 2022 https://arxiv.org/abs/2207.02098

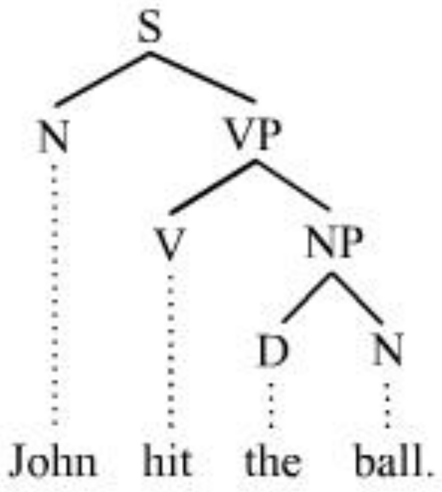
Context-free grammars

- Each production rule is

$$A \rightarrow \alpha$$

where A is a a *single* nonterminal symbol, and α is a string of terminals and/or nonterminals (can be empty)

- CFG define parse trees



Constituency-based parse tree

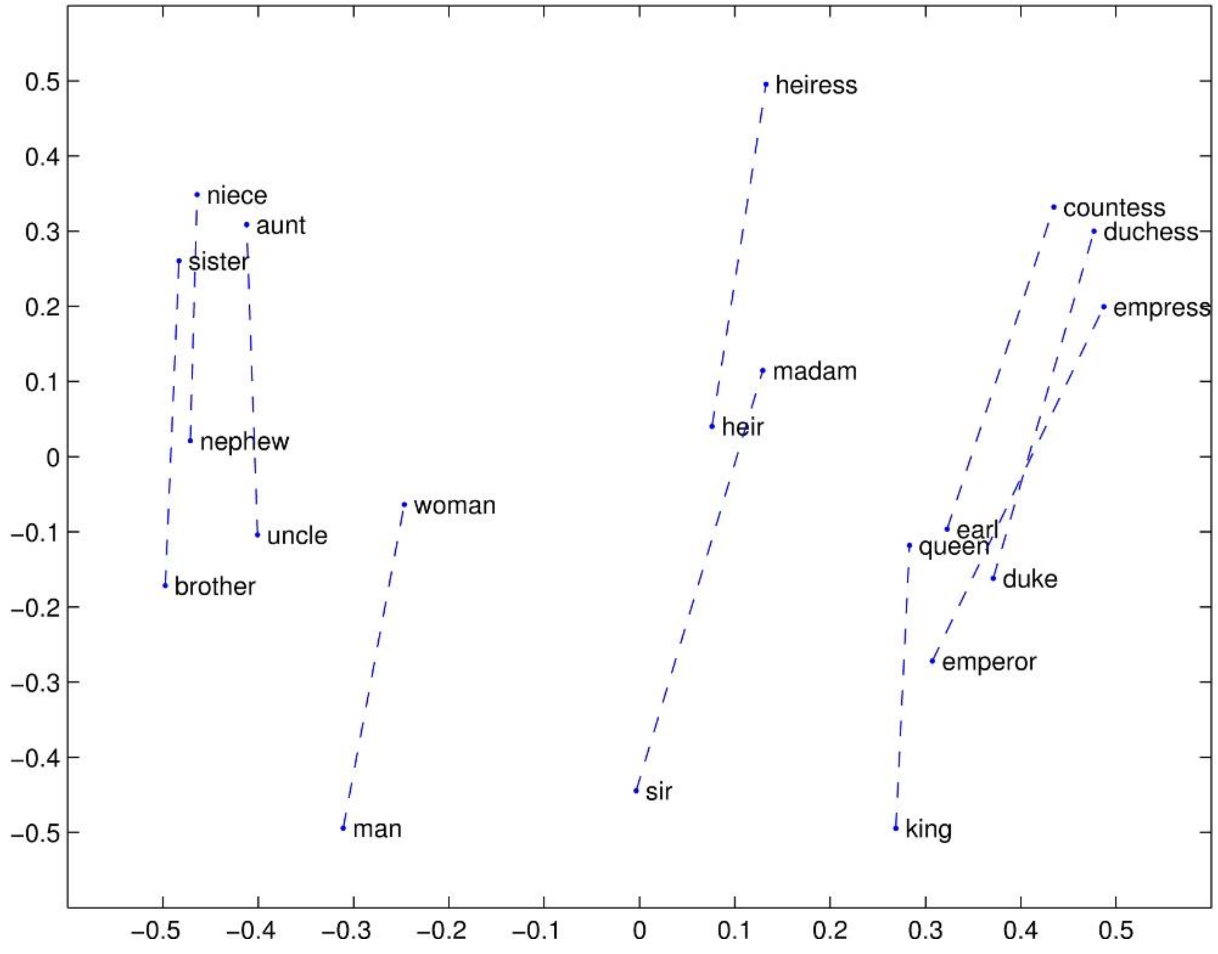
Distributional Hypothesis

- assumes that linguistic items with similar distributions have similar meanings or function
- was likely first introduced by Harris in 1954
- was popularized in the form "a word is characterized by the company it keeps" by Firth
- 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

1. Firth, J.R. A synopsis of linguistic theory 1930-1955 // Studies in Linguistic Analysis, 1957, P. 1-32. Oxford: Philological Society.
2. Harris, Z. Distributional structure // *Word*, 1954, №10(23), P. 146-162.
3. Osgood C., Suci G., Tannenbaum P. The measurement of meaning. — University of Illinois Press, 1957

GloVe

- A (not long ago) popular distributional semantics model
- Comes from two ideas:
    - Distributional hypothesis
    - Word analogy



Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global Vectors for Word Representation]

BERT marked the start of 3rd generation distributional semantic models

- Let’s combine the word and its current context into a single vector
- Let’s select this vector so that, based on the set of these vectors, the missing words in the sentence and the following sentences can be easily guessed

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

Probabilistic language models

- Random sources emits tokens with certain probabilities
- Parameters of a set of sources are determined so that the probability of real texts is higher than "bad" ones.

Марковъ А.А. Примѣръ статистическаго изслѣдованія надъ текстомъ “Евгенія Онѣгина”, иллюстрирующій связь испытаній въ цѣпь // Извѣстія Императорской Академіи Наукъ. VI серія. 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.)
Lalit R. Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2), pages 179–190.

Autoregressive probabilistic language models

- Consider a sequence $t_{1:m} = \{t_1, t_2, \ldots, t_m\}$ from a lexicon $\mathcal{L}$
- An autoregressive language model estimates a probability of the sequence using the chain rule

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

- N-gram model introduces Markovian assumption that the probability depends only on a limited number of predecessors

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

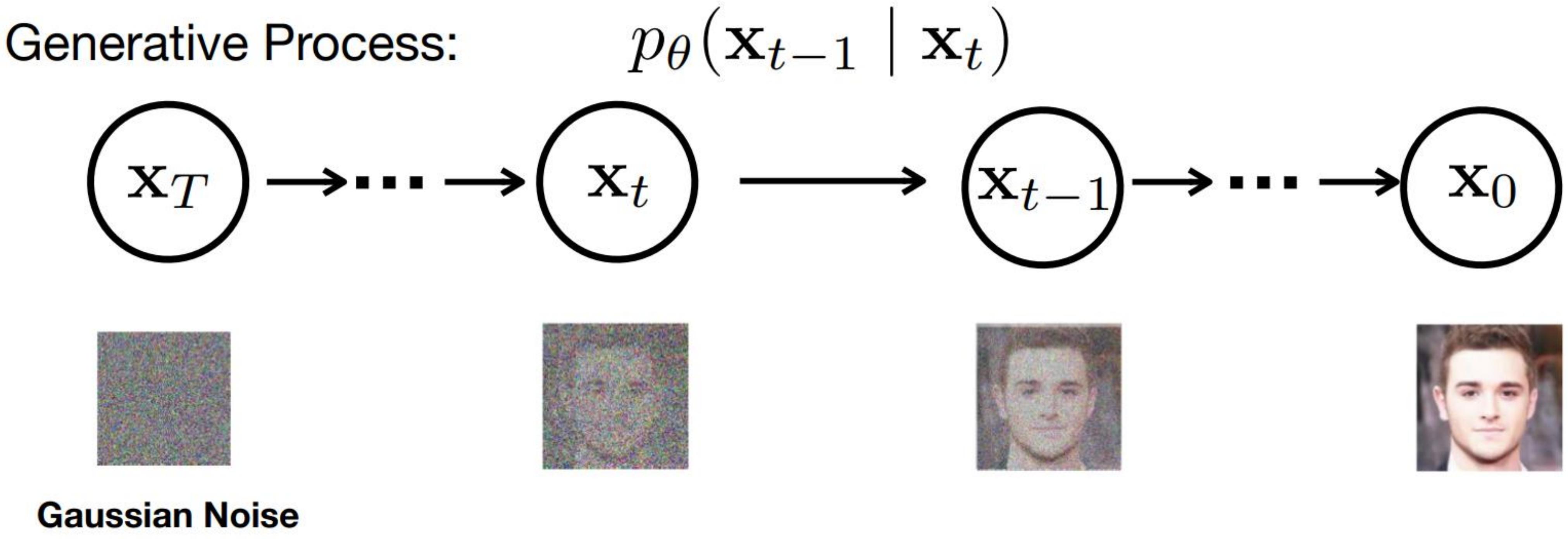
Марковъ А.А. Примѣръ статистическаго изслѣдованія надъ текстомъ “Евгенія Онѣгина”, иллюстрирующій связь испытаній въ цѣпь // Извѣстія Императорской Академіи Наукъ. VI серія. 1913. Vol. 7, № 3. P. 153–162. In Russian. (English translation: Andrei Markov. 2006, An Example of Statistical Investigation of 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.)

Diffusion-LM: Diffusion based Language Models

Diffusion Model for Images is very successful!

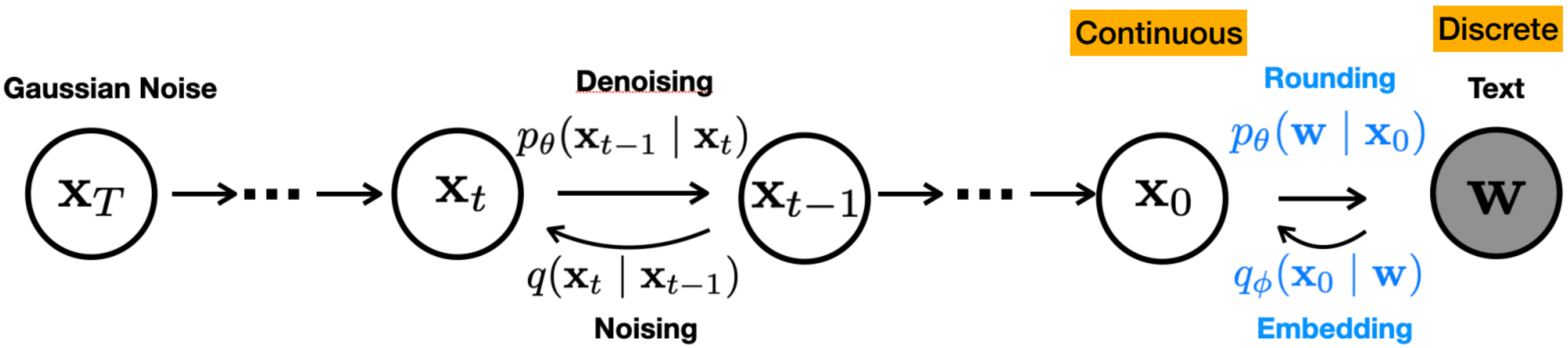
Xiang Lisa Li. Diffusion Models for Text. Beyond autoregressive language modeling. https://web.stanford.edu/class/cs224u/slides/lisa-224u-diffusion.pdf

Diffusion Model for Images



Xiang Lisa Li. Diffusion Models for Text. Beyond autoregressive language modeling.
https://web.stanford.edu/class/cs224u/slides/lisa-224u-diffusion.pdf

Diffusion Model for Discrete Text



Xiang Lisa Li. Diffusion Models for Text. Beyond autoregressive language modeling. https://web.stanford.edu/class/cs224u/slides/lisa-224u-diffusion.pdf

Li, Xiang, John Thickstun, Ishaan Gulrajani, Percy S. Liang, and Tatsunori B. Hashimoto. "Diffusion-LM Improves Controllable Text Generation." *Advances in Neural Information Processing Systems* 35 (2022): 4328-4343.

Why Autocorrelations Decay Laws Matter?

Computing Autocorrelations Using Distributional Semantics

- 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
- cosine distance $d(V_i, V_j) = \cos\angle(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\| \|V_j\|}$

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

- 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.

Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

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

Theorem 1. 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

Theorem 3 . 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

Henry W. Lin and Max Tegmark , Critical Behavior in Physics and Probabilistic Formal Languages, Entropy 2017, 19, 299

Most likely, autoregressive language models exhibit Markovian behavior

- For example, high-quality text generation requires special probabilistic p-sampling (and still degenerates on long text generation).
- Large language models move the border of "long" texts, but do not solve the problem completely

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.

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 world to be a of the…

GPT-2 Continuations
[from Holtzmann et al ICLR 2020]

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg, Pedro A. Ortega. Neural Networks and the Chomsky Hierarchy, 2022 https://arxiv.org/abs/2207.02098

Most likely, autoregressive language models exhibit Markovian behavior

Table 2: Neural architectures and their level in the Chomsky hierarchy as found by our experiments.

Level	Architecture	Description
R-	Transformer [25]	The encoder with stacked multi-head attention layers and dense layers.
R	RNN [23]	A classical RNN with ReLU activations.
R+	LSTM [24]	A classical LSTM.
DCF+	Stack-RNN [30, 32]	An RNN with an external stack, with PUSH, POP, and NO-OP actions.
NDCF	NDStack-RNN [35, 36]	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 [32]).

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Marcus Hutter, Shane Legg, Pedro A. Ortega. Neural Networks and the Chomsky Hierarchy, 2022 https://arxiv.org/abs/2207.02098

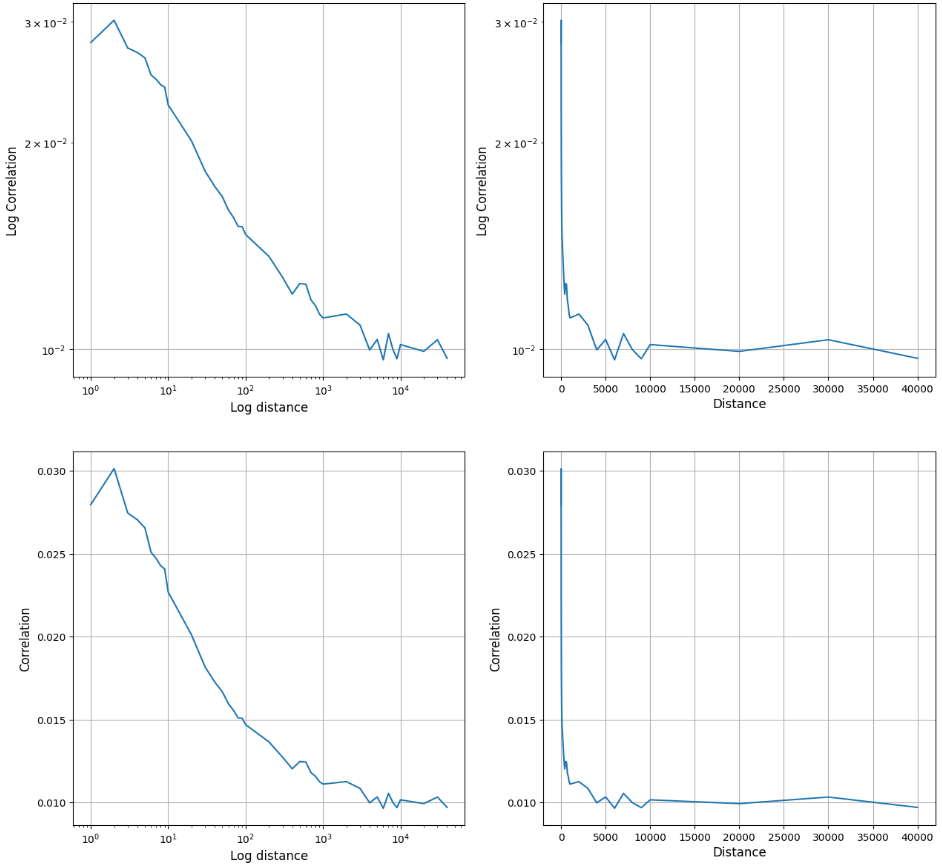
If the Natural Language Exhibits Power Law Correlations Decay, We Can Do Better Than Autoregressive Language Models

Research Questions

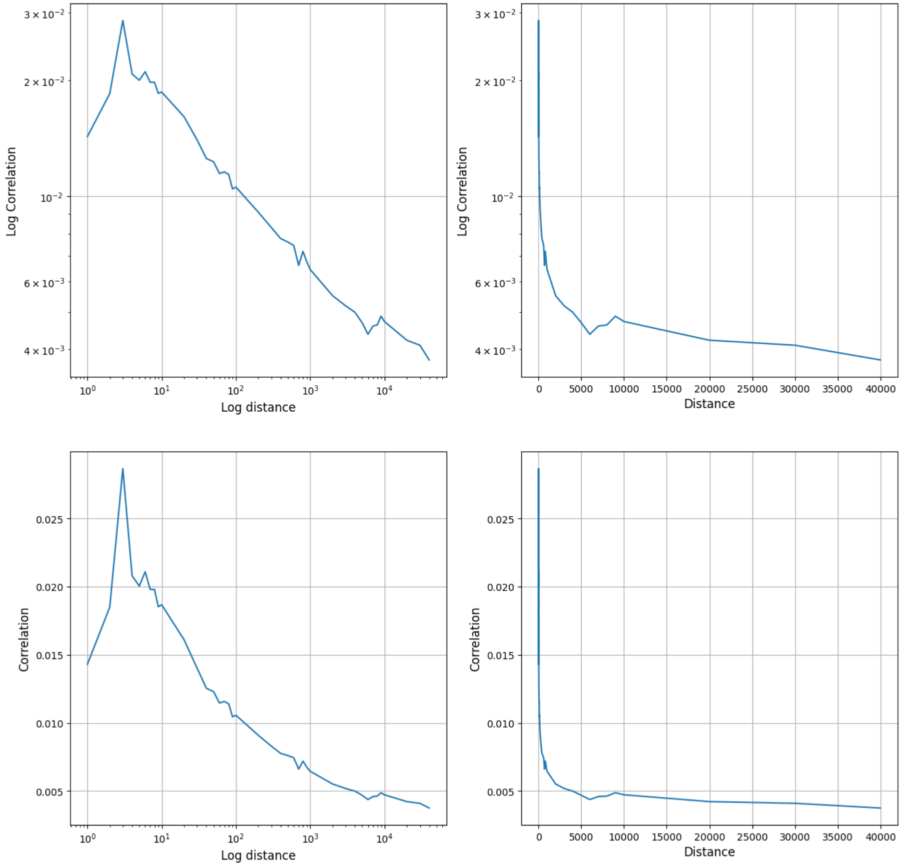
- **Q1.** How accurately can we say that autocorrelations in texts decay according to a power law?
- **Q2.** Can we reject the hypothesis of exponential decay of correlations?
- **Q3.** Does the law of decay depend on the language of the text?
- **Q4.** Over what range of distances does the decay in autocorrelations follow a power law?
- **Q5.** Are autocorrelations in LM-generated texts any different from literary texts?

GloVe Correlations

War and Peace, Russian

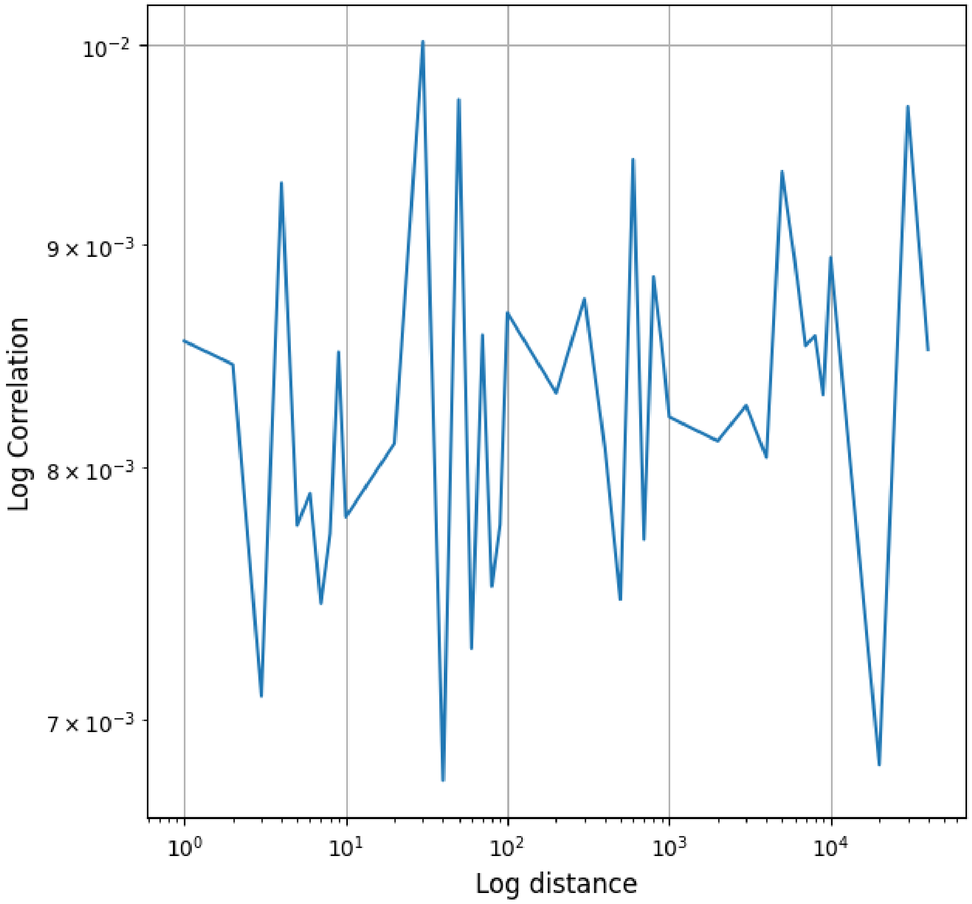


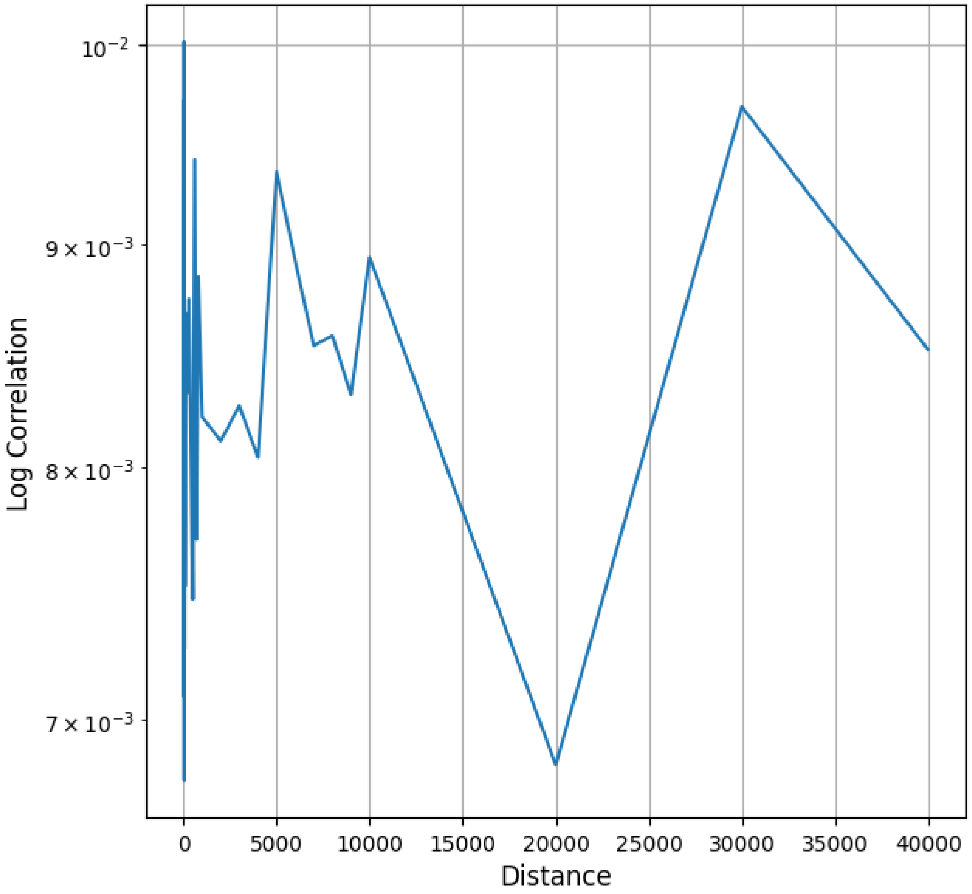
War and Peace, English



Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

Randomly Shuffled Tom Sawyer





Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

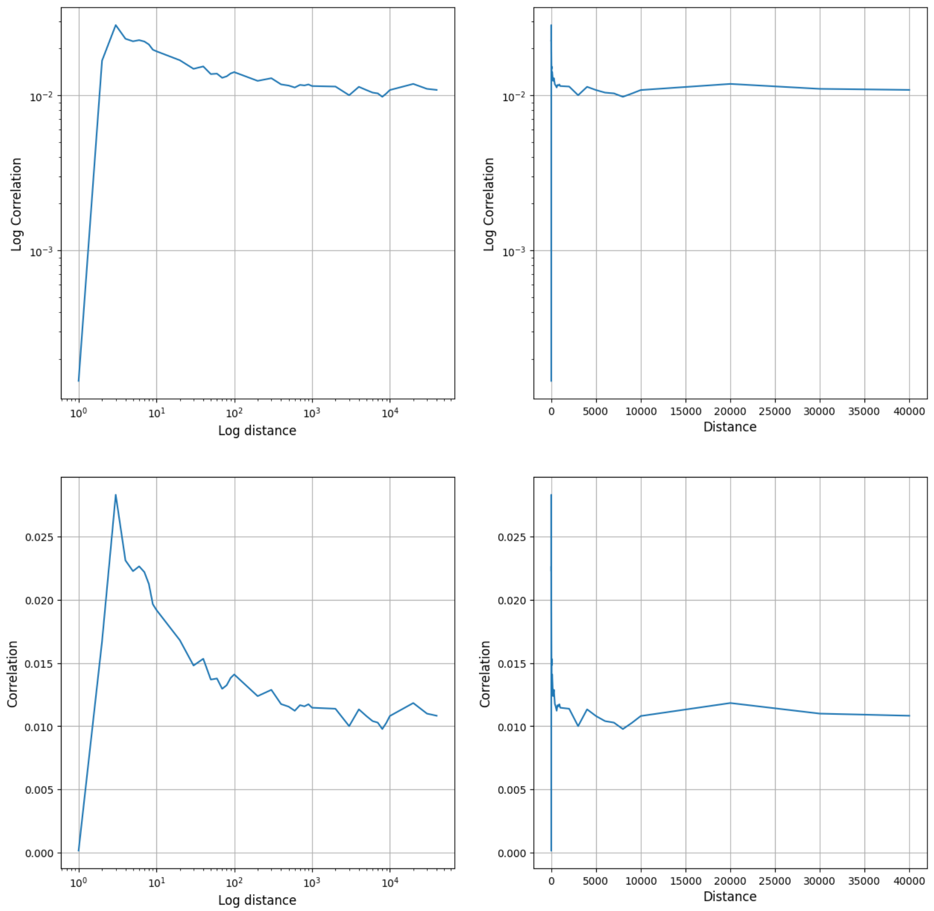
GloVe Correlations Goodness of Fit (MAPE)

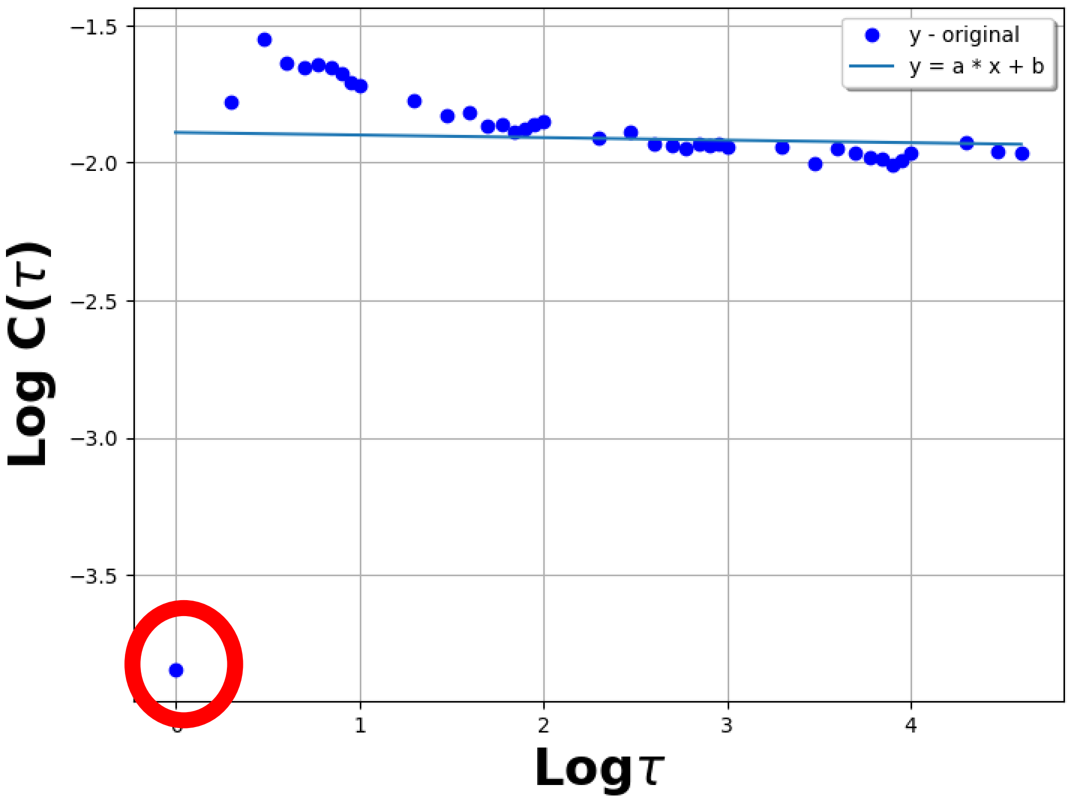
	Power Law					Exponential Law				
	BOW en	fr	es	ru	en	BOW en	fr	es	ru	en
The Adventures of Tom Sawyer	**0,16**	**0,11**	**0,16**	**0,14**	**0,21**	0,52	0,32	0,33	0,33	0,55
The Republic	**0,21**	**0,15**	**0,09**	**0,10**	**0,13**	0,58	0,28	0,25	0,31	0,38
Don Quixote	**0,20**	**0,11**	**0,12**	**0,09**	**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

Mean Absolute Percentage Error

Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

What's Wrong with The Iliad in French?





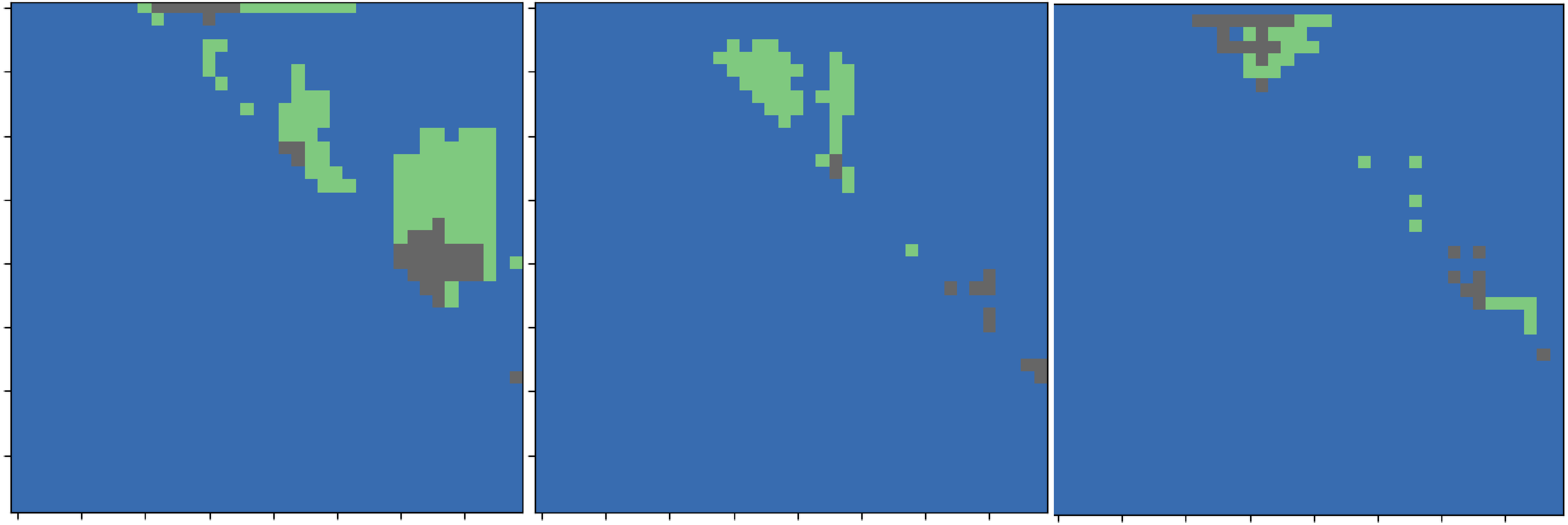
Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

Dependence of the autocorrelations power decay law in Don Quixote on the language and embedding

	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

Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

Dependence of autocorrelations law on distance



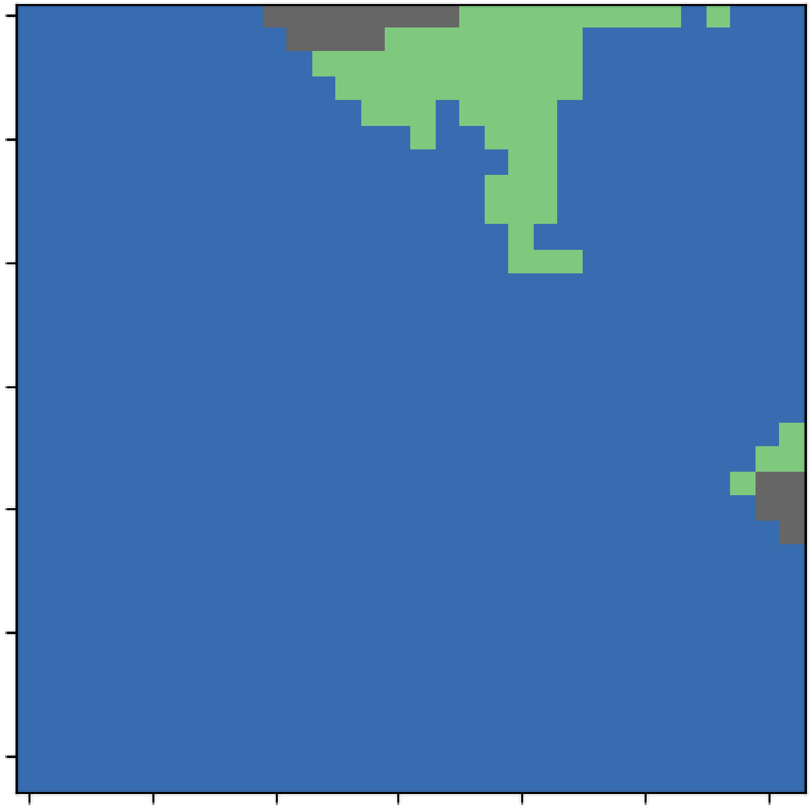
Republic in Spanish   Republic in German   Republic in English

Ranges where power (blue), exp (gray), and log (green) functions are the best approximations

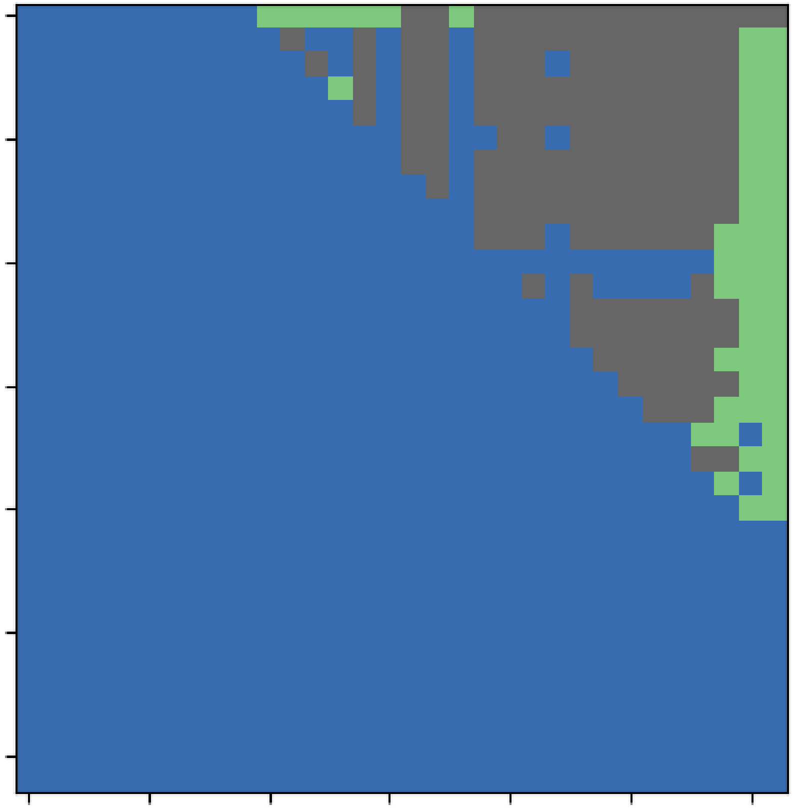
Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

What is the decay law for autocorrelations in LLM-generated texts?

What is the decay law for autocorrelations in LLM-generated texts?



GPT-2



S4

Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

The autocorrelations decay in generated texts is quantitatively different from human texts

- The autocorrelations in generated texts are significantly larger and decay much slower than the ones in the natural texts.
- GPT-2 long texts still suck:

“Humans have always lived on this planet. There has never been a reason that had to change. When a man chooses to go to Africa, why? Because the mosquitoes are very intelligent. They do their job.”

Nikolay Mikhaylovskiy and Ilya Churilov, 2023. Autocorrelations Decay in Texts and Applicability Limits of Language Models. Proceedings of Dialogue-2023

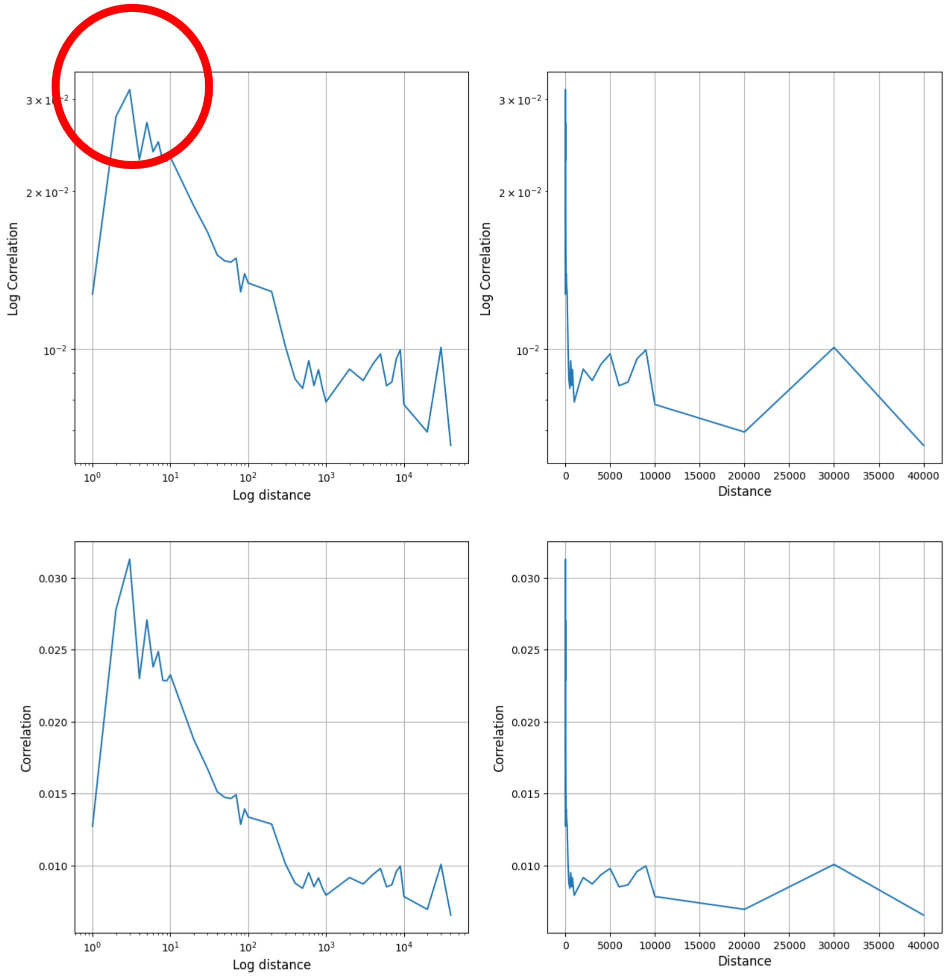
For long text processing one may need architectures different from the autoregressive ones, and many questions remain unanswered.

Immediate research questions

- Is there any dependence of autocorrelation laws on sampling (n-sampling, p-sampling, etc.) temperature and other hyperparameters of GPT-like models and LLAMA zoo?
- How about other models:
  - N-gram
  - RNN
  - LSTM
  - RWKV
  - Diffusion
  - Memory-LSTM
  - … ?

Research Questions: Math and CS

- What is the source of the peak in autocorrelations?
- What is the relationship between different statistical characteristics of texts, such as Zipf's law and the power-law decay of autocorrelation?
- Is structural complexity in the sense of Katznelson related to the applicability of models?



Research Questions: Linguistics

- Do the statistical characteristics of texts depend on:
  - Typologies (agglutinative and synthetic languages, morphosyntactic characteristics, etc.)
  - Type of text (literary, scientific, etc.)
- How does the hierarchical structure implied by generative grammars match the natural hierarchical structures of a long text?
- What operations can be used to extend distribution models?

Research Questions: Engineering

- How to properly use memory in neural network architectures to process long texts?
- How to use the statistical characteristics of texts in the development of quality metrics for applied models?
- Do the statistical characteristics of the text affect the quality of embeddings in transformers?
- What neural network architectures correspond to context-sensitive grammars?
- Is it possible to build language models that explicitly have hierarchical behavior?