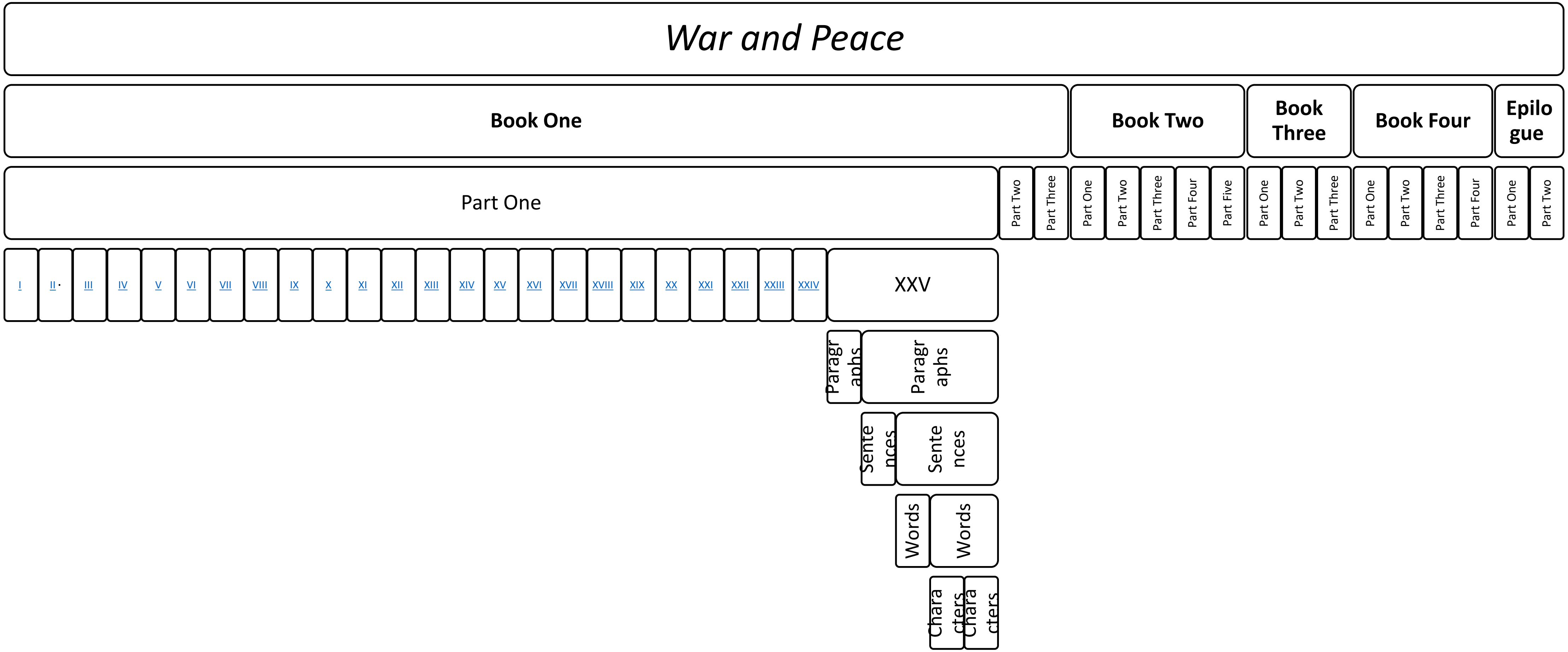
Multiscale Properties of The Language and Language Models

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



Infinite Monkey Theorem

- The infinite monkey theorem states that a monkey hitting keys at **random** on a typewriter keyboard for an infinite amount of time will almost surely type any given text
- We have a random source model
- Once we have a random source model, we can look for correlations in the text. For example, correlations between parts of the same text

- Wait, but what is RANDOM?

Markovian Definition of Random

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

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

- 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.)

The Chomsky Hierarchy

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)

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

Markov models

- describe stochastic regular grammars

PCFG – Probabilistic Context-Free Grammars

- Each production is assigned a probability. The probability of a derivation (parse) is the product of the probabilities of the productions used in that derivation.

R. Thompson and T. Booth, "Applying Probability Measures to Abstract Languages" in IEEE Transactions on Computers, vol. 22, no. 05, pp. 442-450, 1973.

Generation Side of a Random Source Model

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 $I(t1, t2)$ is exponential decay toward zero for $t2 - t1 \gg 1$ with decay timescale log$1\lambda2$, where $\lambda2$ 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

Thus we can observe multiscale, hierarchical structure of a source through autocorrelations

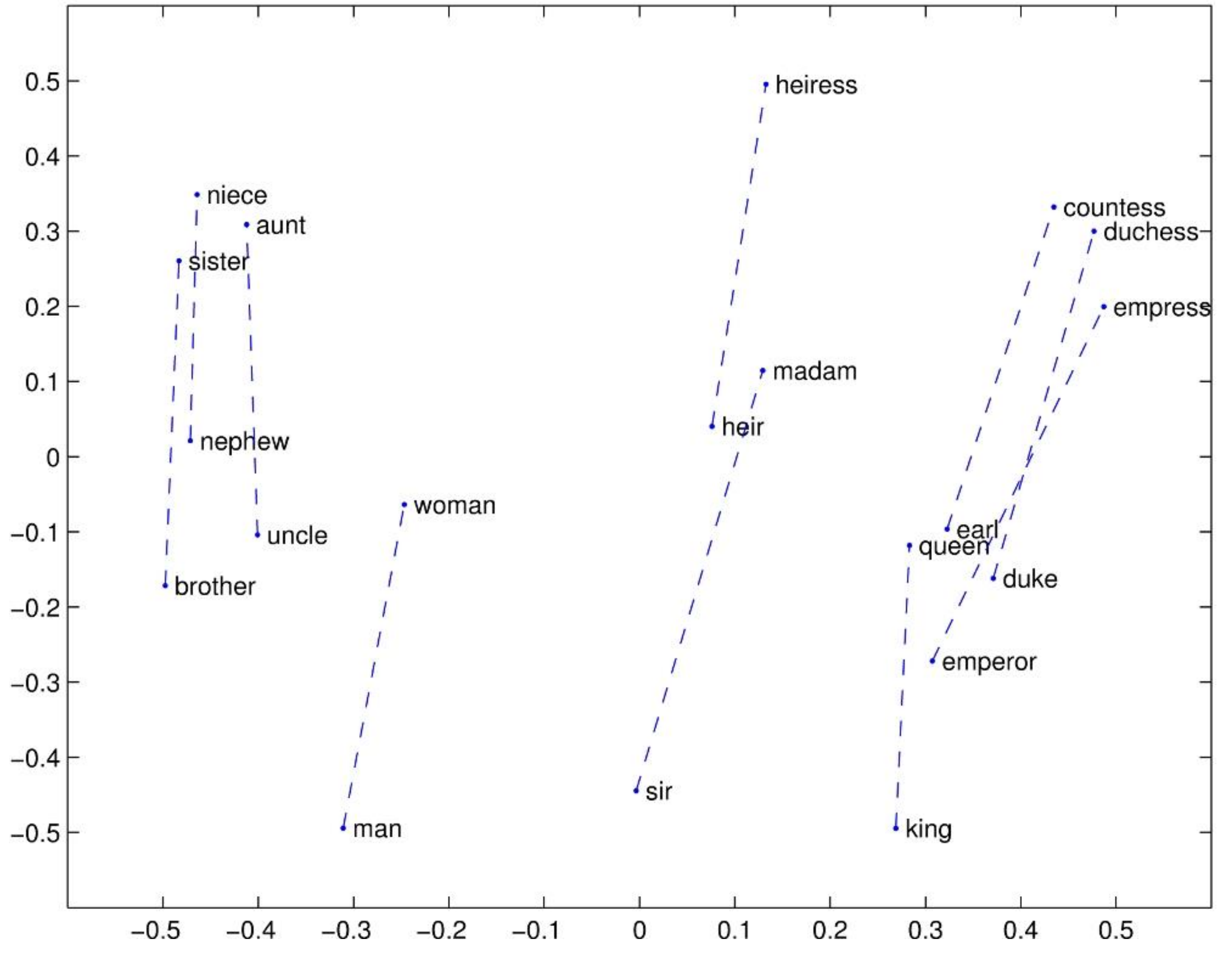
Distributional Hypothesis

- 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.

GloVe

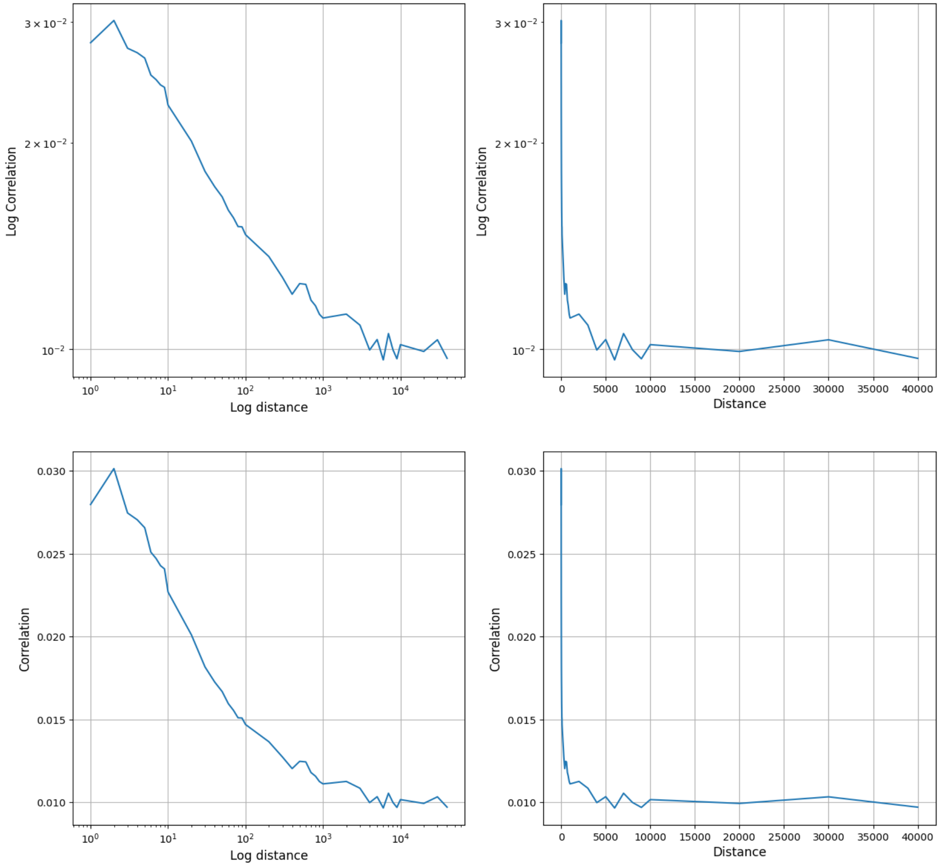
- A (not long ago) popular distributional semantics model
- Comes from two ideas:
  - Distributional hypothesis
  - Word analogy



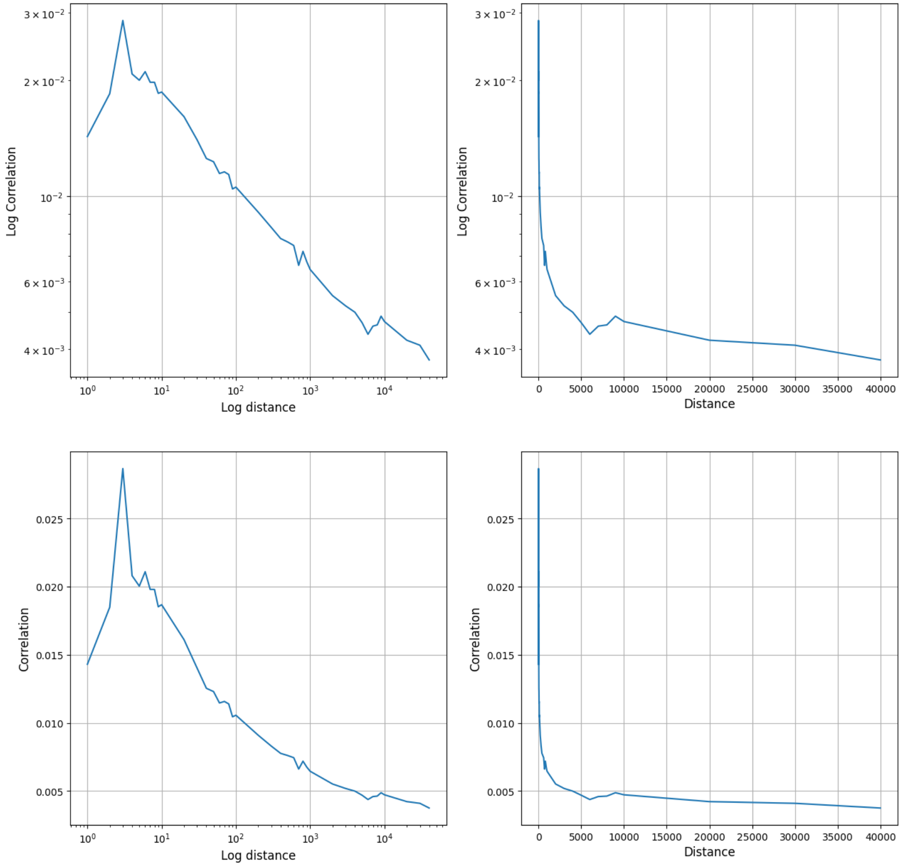
Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global Vectors for Word Representation]

GloVe Correlations

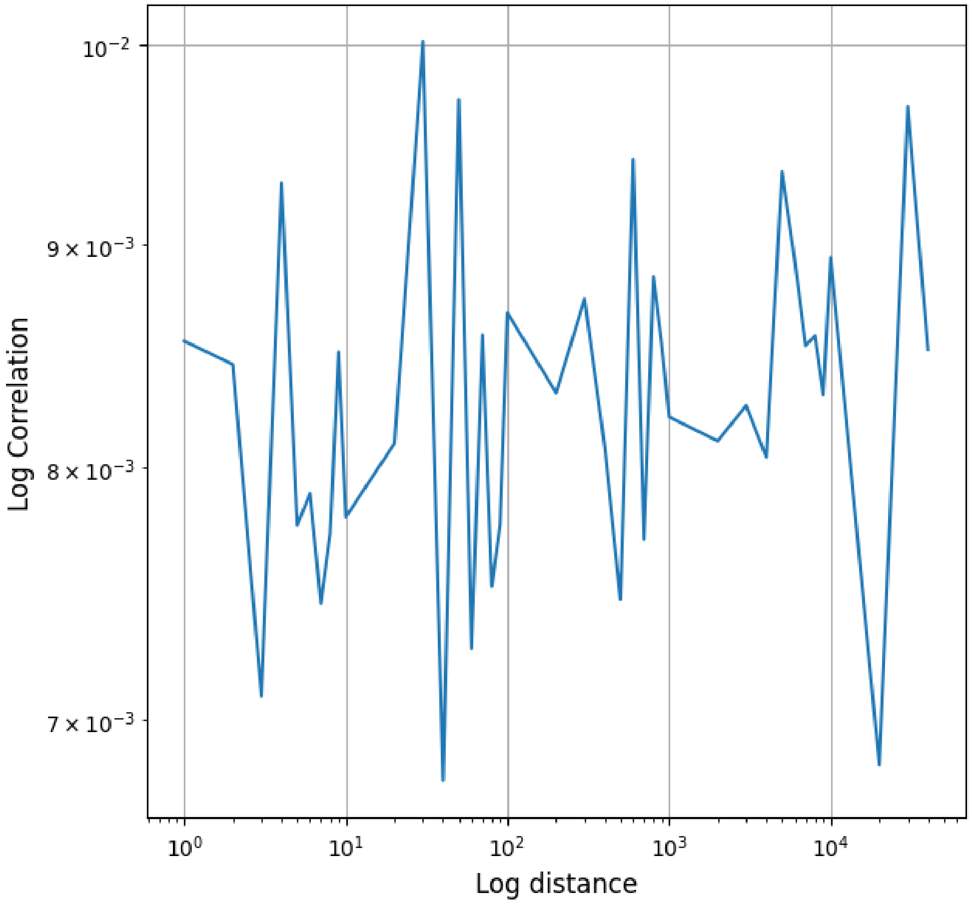
War and Peace, Russian

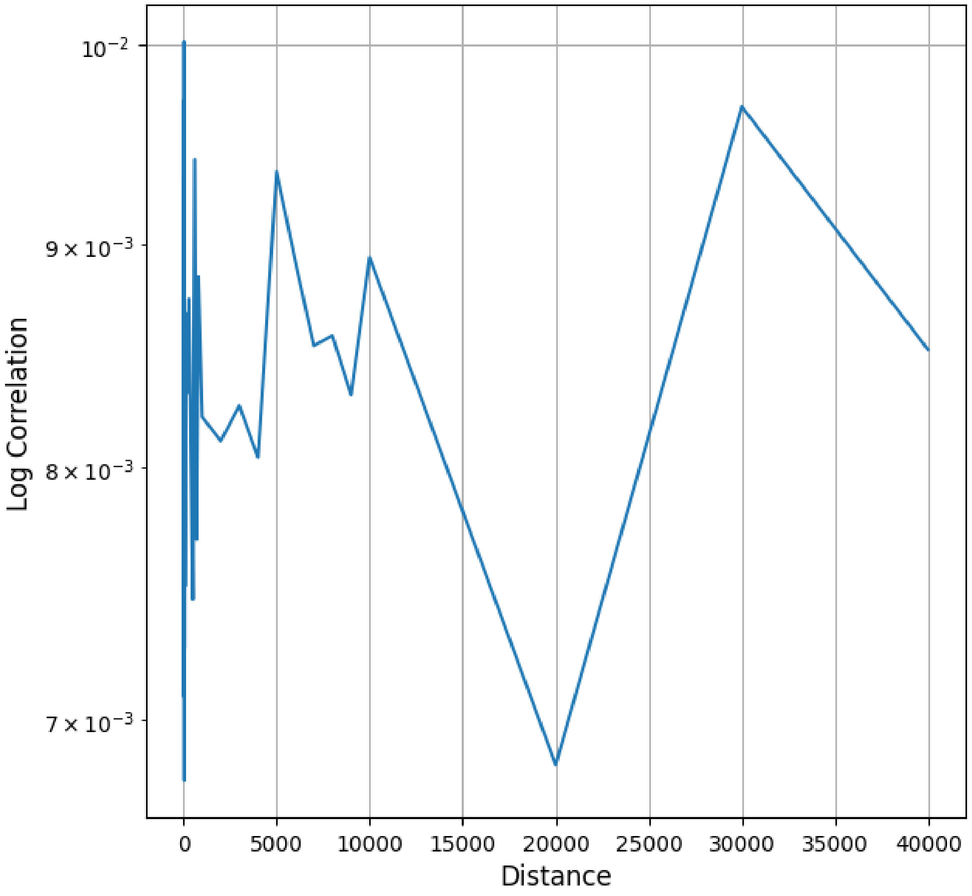


War and Peace, English



Randomly Shuffled Tom Sawyer



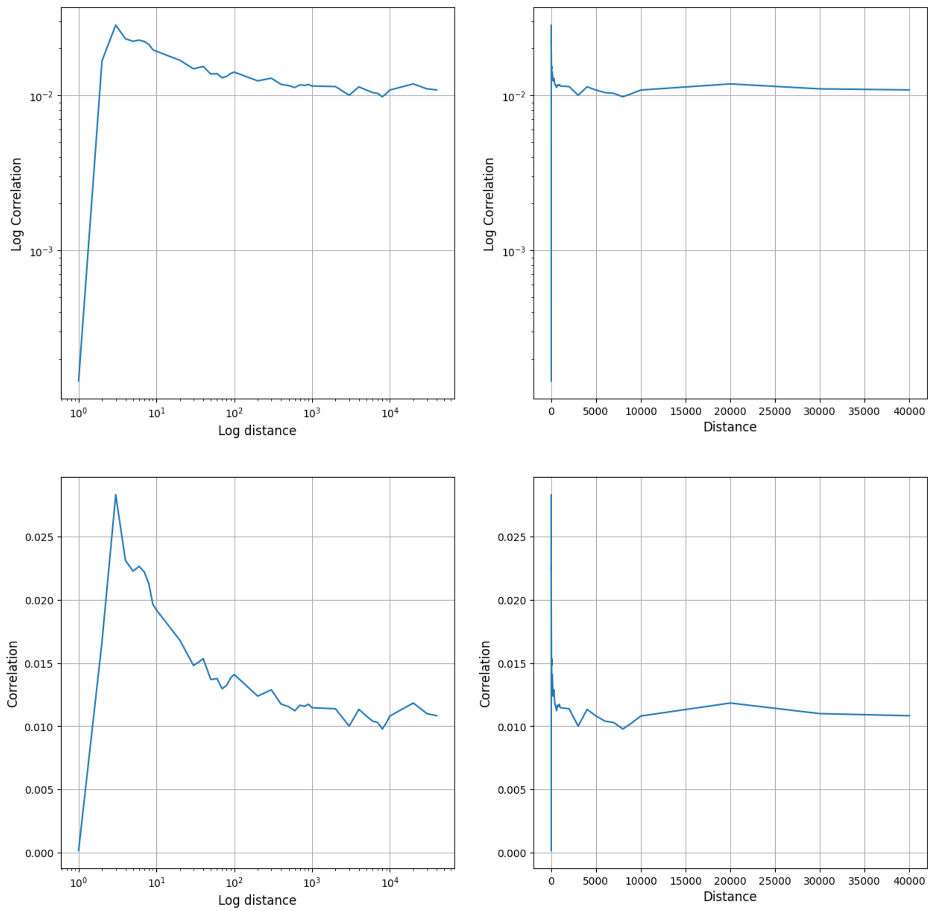


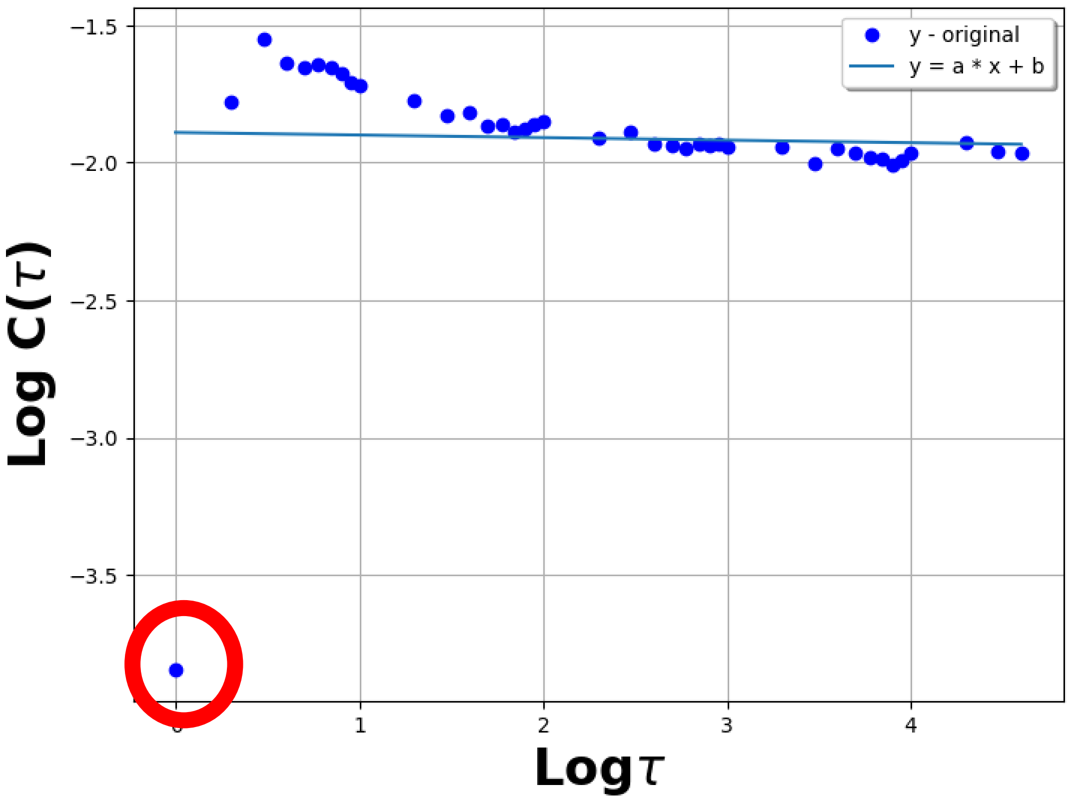
GloVe Correlations Goodness of Fit (MAPE)

	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

What’s Wrong with The Iliad in French?

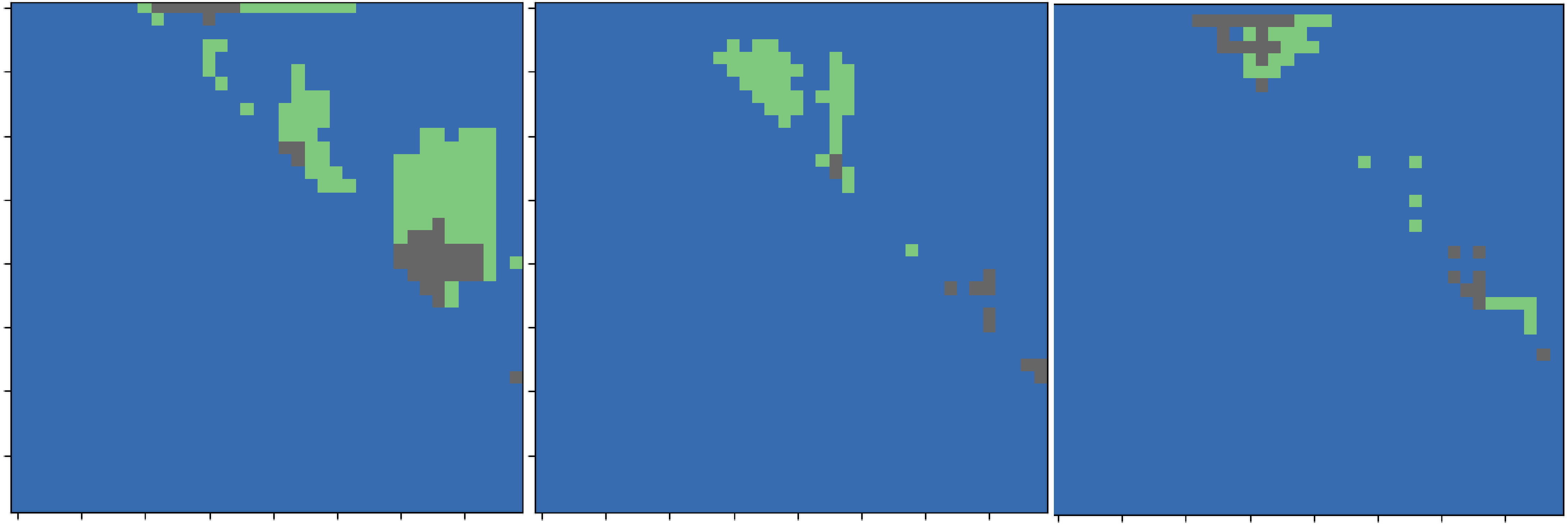




Dependence of the autocorrelations power decay law in Don Quixote on the language and embedding

	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

Dependence of autocorrelations law on distance

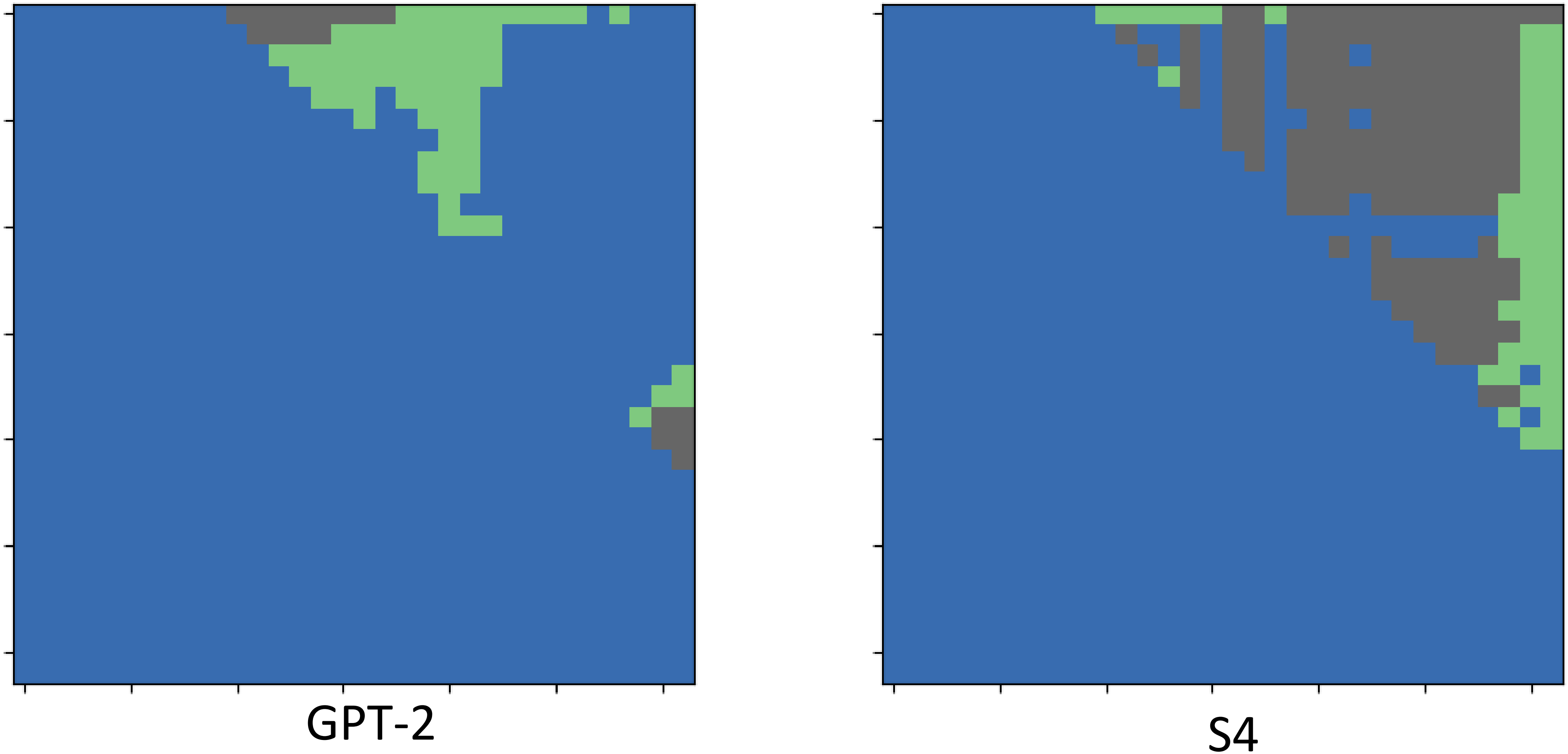


Republic in Spanish | Republic in German | Republic in English

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

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

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



The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts.

We can conclude that for long text processing one may need architectures different from the autoregressive ones, and many questions remain unanswered.