Long-Range Dependence in Word Time Series

Paweł Wieczyński [1] Łukasz Dębowski [2]

[1]no affiliation, Gdańsk, Poland

[2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

QUALICO 2025, Brno, June 26–28

Motivation

- Natural language exhibits complex dependencies beyond a local context
- How far do meaningful correlations extend in texts?
- Current models may miss long-range patterns:
  - n-gram models: Limited by fixed context windows
  - Transformers: Attention mechanisms may be insufficient
- Need to quantify long-range dependence (LRD) in natural language

Long-Range Dependence

Short-Range Dependence (SRD)

Exponential decay of mutual information

$$I(W_i; W_{i+n}) = O(\exp(-\delta n)), \quad \delta > 0$$

Long-Range Dependence (LRD)

Any decay slower than exponential:

- **Power law:** $I(W_i; W_{i+n}) \sim n^{-\gamma}$, $\gamma > 0$
- **Stretched exponential:** $I(W_i; W_{i+n}) \sim \exp(-\delta n^{\beta})$, $0 < \beta < 1$

Implication: LRD indicates non-Markovian behavior

Why not measure Mutual Information directly?

- The Shannon mutual information $I(W_i; W_{i+n})$ is difficult to estimate:
  - Large vocabularies (thousands of words)
  - Strongly dependent sources
  - Sparse data problem
- **Solution:** A computationally tractable proxy measure
- The cosine correlation of word embeddings as a lower bound

Our Approach

Leverage word2vec embeddings to capture semantic relationships while maintaining computational feasibility

Definition of the Cosine Correlation

Cosine Correlation

For random vectors $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\boldsymbol{CC}(\boldsymbol{U};\boldsymbol{V}) := \boldsymbol{E}\left[\frac{\boldsymbol{U}\cdot\boldsymbol{V}}{||\boldsymbol{U}||||\boldsymbol{V}||}\right] - \boldsymbol{E}\left[\frac{\boldsymbol{U}}{||\boldsymbol{U}||}\right]\cdot\boldsymbol{E}\left[\frac{\boldsymbol{V}}{||\boldsymbol{V}||}\right]$$

- Analogy to the Pearson correlation but for vectors
- Properties:
  - $0 \leq \boldsymbol{CC}(\boldsymbol{U};\boldsymbol{U}) \leq 1$
  - $|\boldsymbol{CC}(\boldsymbol{U};\boldsymbol{V})| \leq 1$
  - $\boldsymbol{CC}(\boldsymbol{U};\boldsymbol{V}) = 0$ if $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent

Discrete Case

$$\boldsymbol{CC}(\boldsymbol{U};\boldsymbol{V}) = \sum_{\boldsymbol{u},\boldsymbol{v}} \boldsymbol{\Delta}(\boldsymbol{u},\boldsymbol{v})\,\boldsymbol{\cos}(\boldsymbol{u};\boldsymbol{v})$$

where $\boldsymbol{\Delta}(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{P}(\boldsymbol{U}=\boldsymbol{u},\boldsymbol{V}=\boldsymbol{v}) - \boldsymbol{P}(\boldsymbol{U}=\boldsymbol{u})\boldsymbol{P}(\boldsymbol{V}=\boldsymbol{v})$

Key theoretical result

Theorem (Lower bound via Pinsker inequality)

$$I(U;V) \geq \frac{CC(U;V)^2}{2}$$

Proof sketch

- The Pinsker inequality: $D_{KL}(p||q) \geq \frac{1}{2}\sum_x |p(x) - q(x)|^2$
- Use the Cauchy-Schwarz inequality: $|\cos(u;v)| \leq 1$
- Combine with the definition of the cosine correlation

Consequence

A slower than exponential decay of the cosine correlation

$\Longrightarrow$ Long-range dependence

Thus, we may detect non-Markovian sources through embeddings!

Experimental design

Word Embeddings

- 100-dimensional word2vec from NLPL repository
- Trained on CoNLL17 corpora using continuous skipgram
- Uniform baseline across 17 languages

Pooled Embeddings

$$F_i^{(k)} = \sum_{j=0}^{k-1} F_{i+j}$$

- Pooling orders: $k \in \{1, 3, 9, 27\}$
- Capture the local semantic context
- $k = 1$: individual word embeddings
- Measurement: $C(n|k) = |CC(F_i^{(k)}; F_{i+n}^{(k)})|$

Data sources

Natural Texts: The Standardized Project Gutenberg Corpus

- 17 languages, ∼100 texts each
- Literary texts representing human language usage
- Coverage: 0.92 ± 0.17 of tokens have embeddings
- Text lengths: 38312 ± 34346 tokens

Artificial Texts: The Human vs. LLM Text Corpus

- 1000 human texts + 6000 LLM texts (English only)
- LLM sources:
  GPT-3.5, GPT-4, LLaMA variants (7B, 13B, 30B, 65B)
- Coverage: 0.999 ± 0.005 of tokens have embeddings
- Text lengths:
  492 ± 424 tokens (much shorter than natural texts)

Fitting procedure

Two decay models

Power law: $f(n|c,\gamma) = cn^{-\gamma}$

Stretched exponential: $f(n|b,\delta,\beta) = \exp(-\delta n^{\beta} + b)$

Methodology

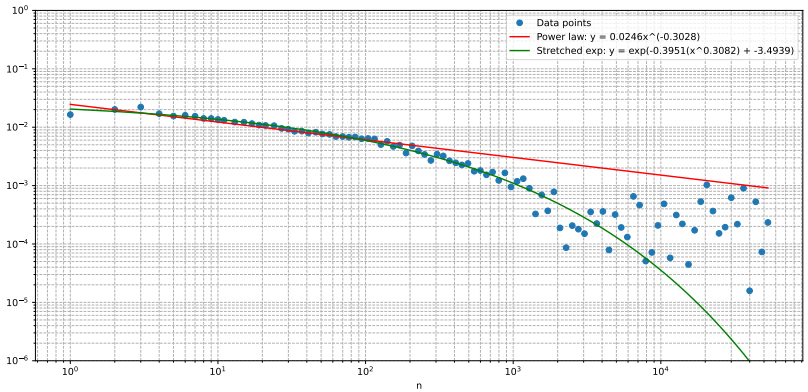
- Fitting range: $k \leq n \leq 1000$ words
- Goodness-of-fit metric:
  Sum of Squared Logarithmic Residuals (SSLR)
- Parameter estimation:
  SciPy curve_fit with trust region reflective
- Statistical testing: Kruskal-Wallis test for parameter distributions, post-hoc Dunn test

Key insight: Signal dissolves into noise around $n = 1000$ words

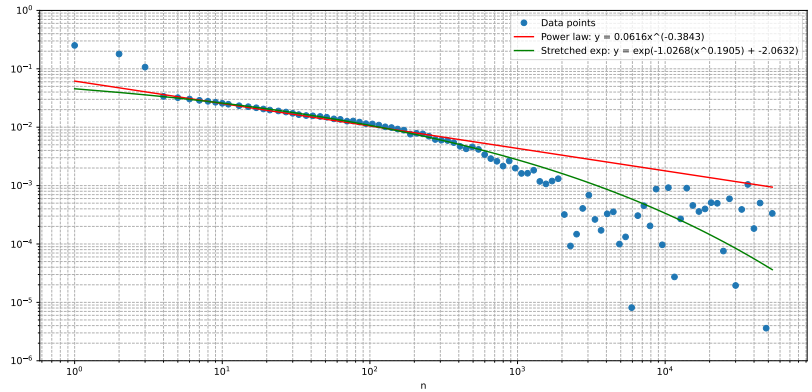
Cosine correlation estimates $C(n|k)$

Cecilia: A Story of Modern Rome in English:

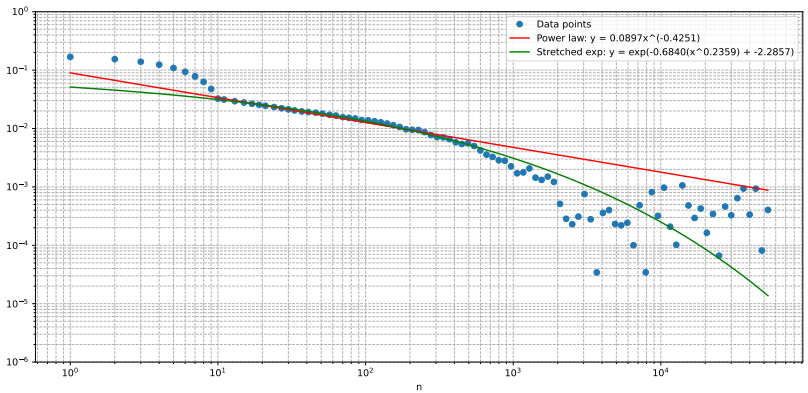
$k = 1$:



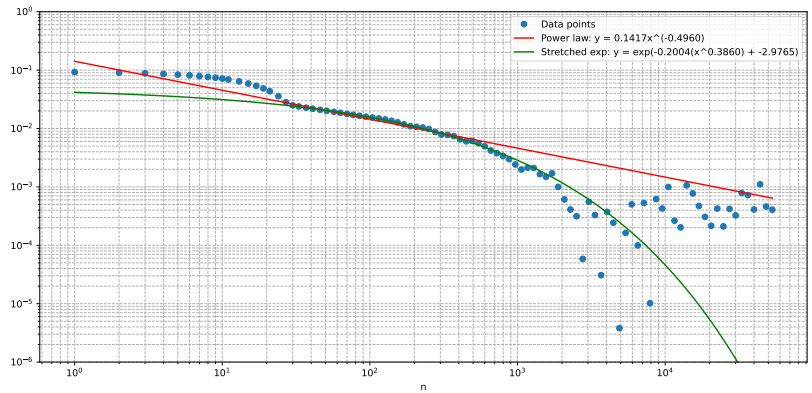
$k = 3$:



$k = 9$:



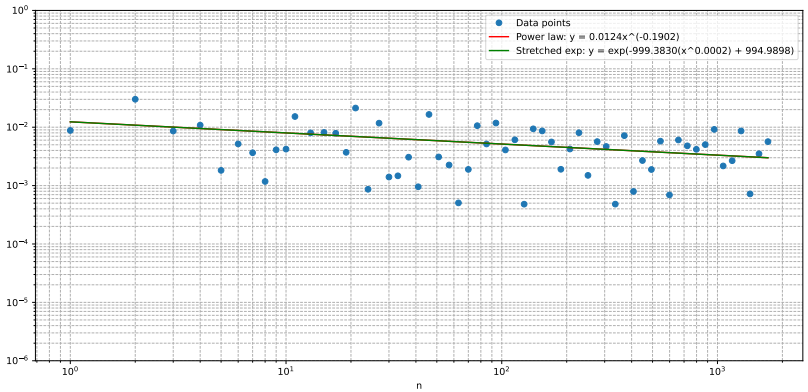
$k = 27$:



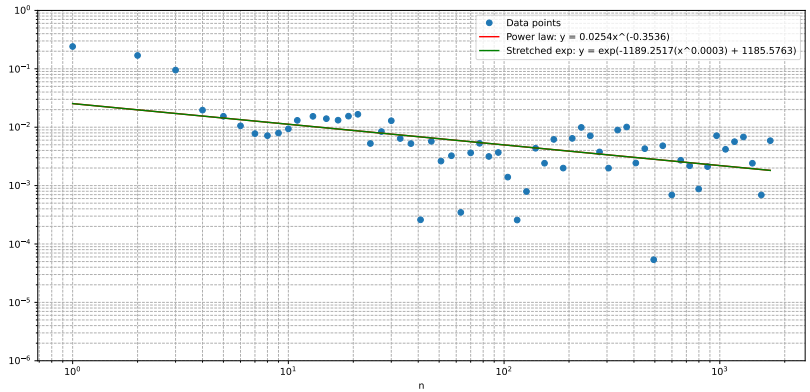
Cosine correlation estimates $C(n|k)$

Text no. 702 by GPT 3.5:

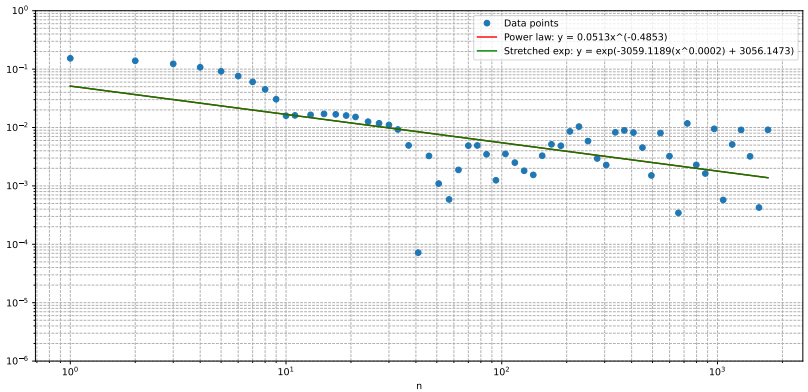
$k = 1$:



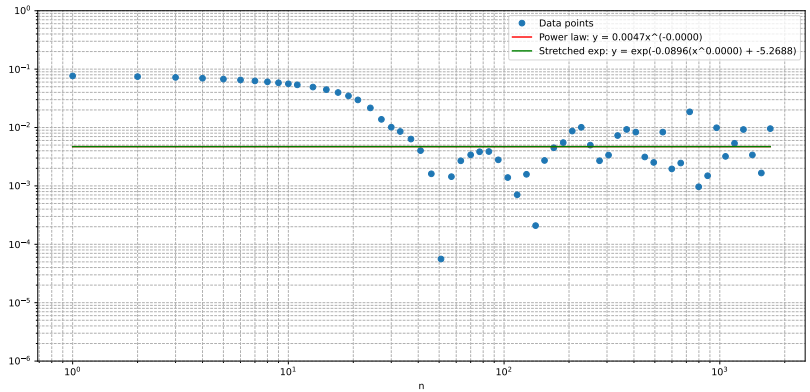
$k = 3$:



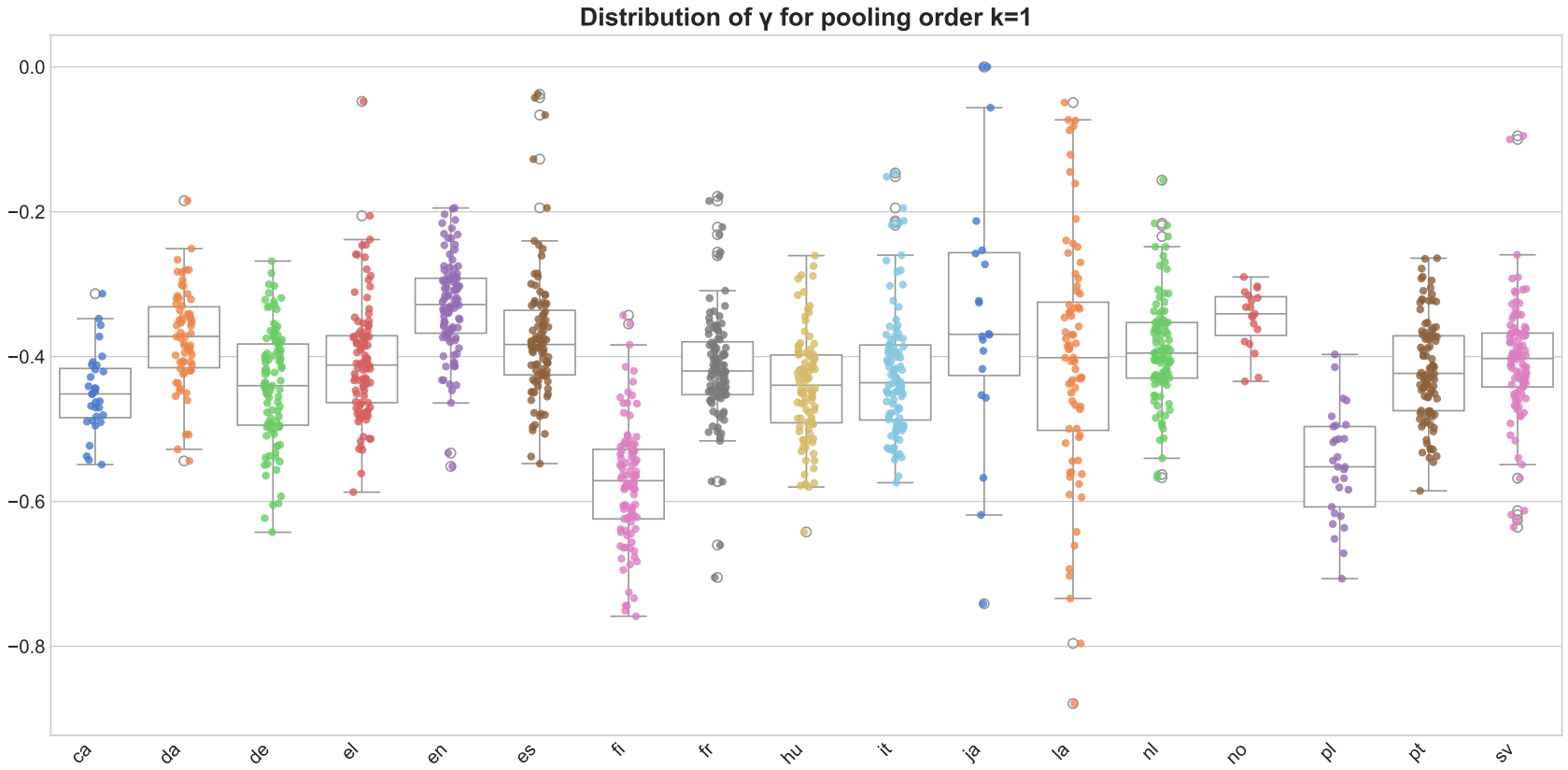
$k = 9$:



$k = 27$:



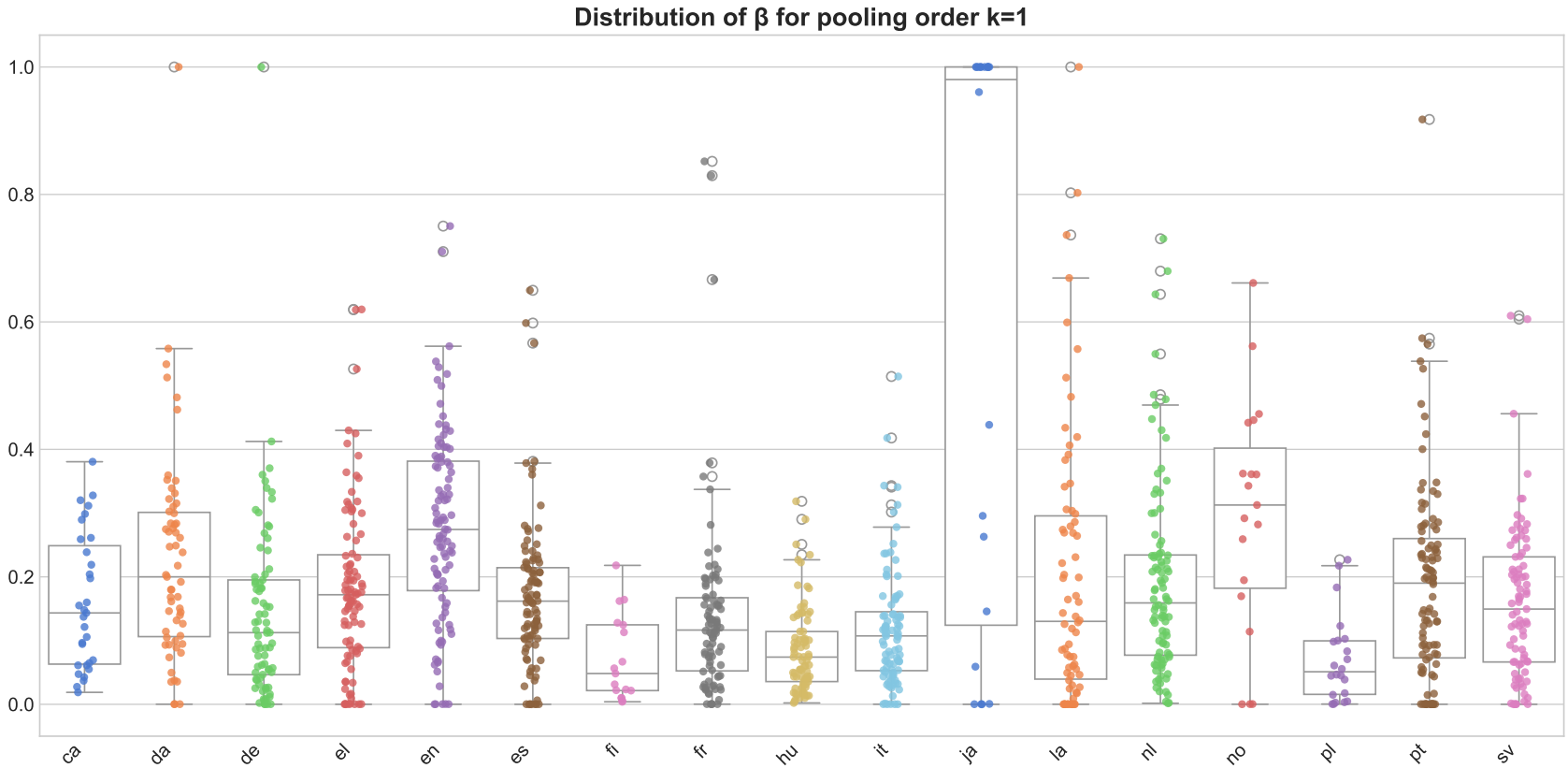
The power law exponent γ for $k = 1$



The power law exponent γ

	$k = 1$	$k = 3$	$k = 9$	$k = 27$
ca	0.449 ± 0.055	0.523 ± 0.063	0.546 ± 0.083	0.62 ± 0.13
da	0.373 ± 0.067	0.442 ± 0.092	0.49 ± 0.11	0.58 ± 0.17
de	0.440 ± 0.079	0.49 ± 0.11	0.53 ± 0.14	0.62 ± 0.23
el	0.405 ± 0.083	0.48 ± 0.13	0.51 ± 0.16	0.57 ± 0.23
en	0.330 ± 0.067	0.418 ± 0.098	0.47 ± 0.12	0.54 ± 0.19
es	0.373 ± 0.090	0.44 ± 0.11	0.45 ± 0.12	0.50 ± 0.16
fi	0.574 ± 0.084	0.552 ± 0.099	0.57 ± 0.13	0.62 ± 0.17
fr	0.415 ± 0.079	0.47 ± 0.11	0.49 ± 0.13	0.55 ± 0.20
hu	0.442 ± 0.075	0.482 ± 0.096	0.49 ± 0.11	0.55 ± 0.17
it	0.421 ± 0.089	0.47 ± 0.12	0.48 ± 0.13	0.52 ± 0.19
ja	0.34 ± 0.19	0.57 ± 0.39	0.79 ± 0.69	2.3 ± 1.6
la	0.40 ± 0.17	0.47 ± 0.23	0.51 ± 0.24	0.60 ± 0.33
nl	0.390 ± 0.074	0.46 ± 0.11	0.51 ± 0.13	0.60 ± 0.20
no	0.347 ± 0.041	0.451 ± 0.060	0.522 ± 0.073	0.59 ± 0.10
pl	0.550 ± 0.075	0.63 ± 0.12	0.65 ± 0.18	0.73 ± 0.26
pt	0.416 ± 0.071	0.57 ± 0.16	0.63 ± 0.22	0.70 ± 0.47
sv	0.407 ± 0.084	0.459 ± 0.095	0.51 ± 0.12	0.60 ± 0.17
GPT-3.5	0.31 ± 0.36	0.19 ± 0.36	0.9 ± 1.9	1.2 ± 2.0
GPT-4	0.47 ± 0.30	0.39 ± 0.42	0.6 ± 1.3	0.8 ± 1.6
Human	0.28 ± 0.26	0.29 ± 0.38	0.7 ± 1.3	0.9 ± 1.3
LLaMA-13B	0.19 ± 0.20	0.25 ± 0.27	0.6 ± 1.3	0.8 ± 1.4
LLaMA-30B	0.20 ± 0.20	0.25 ± 0.35	0.6 ± 1.3	0.8 ± 1.3
LLaMA-65B	0.19 ± 0.20	0.23 ± 0.27	0.6 ± 1.2	0.8 ± 1.2
LLaMA-7B	0.21 ± 0.20	0.25 ± 0.24	0.6 ± 1.0	0.7 ± 1.1

The stretched exponential exponent β for $k = 1$



The stretched exponential exponent β

	$k = 1$	$k = 3$	$k = 9$	$k = 27$
ca	0.16 ± 0.11	0.090 ± 0.093	0.16 ± 0.15	0.24 ± 0.21
da	0.23 ± 0.17	0.17 ± 0.11	0.21 ± 0.16	0.23 ± 0.19
de	0.14 ± 0.15	0.14 ± 0.12	0.18 ± 0.15	0.27 ± 0.19
el	0.18 ± 0.13	0.14 ± 0.15	0.18 ± 0.17	0.30 ± 0.24
en	0.28 ± 0.15	0.17 ± 0.14	0.21 ± 0.15	0.31 ± 0.23
es	0.17 ± 0.12	0.11 ± 0.15	0.15 ± 0.15	0.23 ± 0.20
fi	0.071 ± 0.067	0.067 ± 0.065	0.12 ± 0.11	0.21 ± 0.15
fr	0.14 ± 0.14	0.13 ± 0.15	0.16 ± 0.19	0.24 ± 0.23
hu	0.086 ± 0.069	0.093 ± 0.090	0.18 ± 0.16	0.26 ± 0.25
it	0.119 ± 0.095	0.10 ± 0.13	0.13 ± 0.15	0.20 ± 0.18
ja	0.61 ± 0.45	0.55 ± 0.49	0.59 ± 0.48	0.49 ± 0.48
la	0.20 ± 0.22	0.32 ± 0.25	0.40 ± 0.30	0.52 ± 0.37
nl	0.19 ± 0.15	0.15 ± 0.13	0.19 ± 0.15	0.27 ± 0.21
no	0.30 ± 0.18	0.22 ± 0.11	0.187 ± 0.081	0.23 ± 0.18
pl	0.070 ± 0.068	0.090 ± 0.089	0.15 ± 0.17	0.24 ± 0.24
pt	0.19 ± 0.16	0.13 ± 0.16	0.18 ± 0.20	0.23 ± 0.25
sv	0.16 ± 0.12	0.17 ± 0.13	0.21 ± 0.15	0.29 ± 0.24
GPT-3.5	0.07 ± 0.24	0.11 ± 0.30	0.18 ± 0.37	0.28 ± 0.44
GPT-4	0.02 ± 0.12	0.04 ± 0.19	0.06 ± 0.22	0.11 ± 0.30
Human	0.10 ± 0.27	0.19 ± 0.35	0.27 ± 0.41	0.32 ± 0.44
LLaMA-13B	0.24 ± 0.38	0.26 ± 0.39	0.25 ± 0.40	0.32 ± 0.45
LLaMA-30B	0.21 ± 0.36	0.24 ± 0.38	0.24 ± 0.40	0.31 ± 0.44
LLaMA-65B	0.20 ± 0.36	0.23 ± 0.38	0.26 ± 0.42	0.33 ± 0.45
LLaMA-7B	0.25 ± 0.39	0.27 ± 0.39	0.27 ± 0.41	0.30 ± 0.44

Conclusion

- ✓ **Natural language exhibits LRD up to 1000 words**
  - 4 decades larger than for character-level studies
  - Systematic across multiple languages
- ✓ **Stretched exponential better describes natural texts**
  - More stable fits than the power law
  - Language-specific parameters
- × **LLM-generated texts lack systematic LRD patterns**
  - Dominated by noise
  - Poor goodness of fit across all LLM types

Core insight

Cosine correlation provides a practical tool for detecting long-range dependence in natural language

Limitations of current LLMs

Empirical evidence

- GPT and LLaMA variants fail to exhibit systematic LRD
- High variance in the fitted parameters
- Poor goodness of fit across all models

Possible explanations

- **Text length:** LLM texts are shorter (492 vs. 38312 tokens)
- **Training objective:** The next-token prediction may not capture long-range structure
- **Architecture:** Transformer attention may be fundamentally limited

The gap between natural and artificial text generation reveals limitations in the current AI approaches to language modeling

The full paper

- **Article:**
  https://www.mdpi.com/1099-4300/27/6/613
- **Code Repository:**
  https://github.com/pawel-wieczynski/long_range_dependencies