Menzerath-Altmann's law versus Menzerath's law as a criterion of complexity in communication

Iván G. Torre [1,2] Łukasz Dębowski [3]
Antonio Hernández-Fernández [4,5]

[1] Language and Speech Laboratory,
Universidad del País Vasco/Euskal Herriko Unibertsitatea

[2] Departamento de Matemática Aplicada, Universidad Politécnica de Madrid

[3] Institute of Computer Science, Polish Academy of Sciences

[4] Societat Catalana de Tecnologia, Secció de Ciències i Tecnologia,
Institut d'Estudis Catalans

[5] Institut de Ciències de l'Educació, Universitat Politècnica de Catalunya

QUALICO 2023, Lausanne, June 28–30

Introduction

A terminological distinction:

- **Menzerath's law** (ML) is a QL law which states that, on average, the longer is a linguistic construct, the shorter are its constituents.
- **Menzerath-Altmann's law** (MAL) is a precise mathematical formula which expresses the expected length of the linguistic construct conditioned on the number of its constituents.

Our contribution:

- We derive ML for *monkey typing* with three emitted symbols. Thus the general ML doesn't distinguish language from *monkey-typing*! It is only the exact form of MAL that does!
- We analyze MAL on data from 21 languages, consisting of texts from the Standardized Project Gutenberg. We report an inverted regime, not exhibited by *monkey-typing*.

Syllables—how to count them?

Syllable is a structure made up of a sound sequence with:

- an optional *onset* (a consonant or a consonant cluster),
- a compulsory *nucleus* (commonly a vowel or a syllabic consonant),
- an optional *coda* (a consonant or a consonant cluster).

The expected length of a syllable:

- $\boldsymbol{N}$ — the number of consonants in a word;
- $\boldsymbol{M}$ — the number of vowels in a word;
- $\frac{\boldsymbol{N}+\boldsymbol{M}}{\boldsymbol{M}}$ — the mean length of a syllable in a word.

Menzerath's and Menzerath-Altmann's laws

The expected length of a syllable in an m-syllable word:

$$\mathbb{E}\left(\frac{N+M}{M}\middle|M=m\right)=\sum_{n=0}^{\infty}\frac{n+m}{m}\cdot\frac{P(N=n,M=m)}{P(M=m)}.$$

Menzerath's law (1928):

$$\mathbb{E}\left(\frac{N+M}{M}\middle|M=m+1\right)\leq\mathbb{E}\left(\frac{N+M}{M}\middle|M=m\right),\quad m\in\mathbb{N}.$$

Menzerath-Altmann's law (1980):

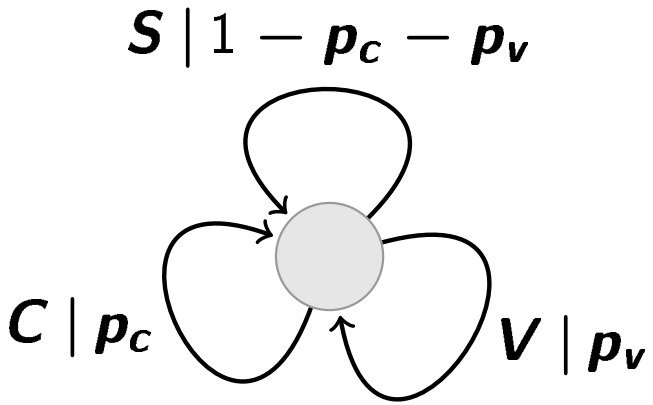
$$\mathbb{E}\left(\frac{N+M}{M}\middle|M=m\right)=\alpha m^{\beta}\exp(-\gamma m),$$

If $\beta\cdot\gamma>0$ then there is an extremum at $n=\beta/\gamma$ and there is an inverted regime in MAL (Torre et al. 2019).

A monkey-typing model of Menzerath's law

Consider a memoryless source that emits symbols:

- C (consonant) with probability p_c,
- V (vowel) with probability p_v,
- S (space) with probability $1 - p_c - p_v$.



We define:

- N — the number of C's generated between two S's,
- M — the number of V's generated between two S's.

The Menzerath law for monkey typing

Let N_i — the number of consecutive C's generated directly after the i-th V (or after S in case of $i = 0$). Then may write:

$$\mathbb{E}\,(N|M = m) = \sum_{i=0}^{m} \mathbb{E}\,(N_i|M = m) = \sum_{i=0}^{m} \mathbb{E}\,N_i = (m+1)\,\mathbb{E}\,N_0$$

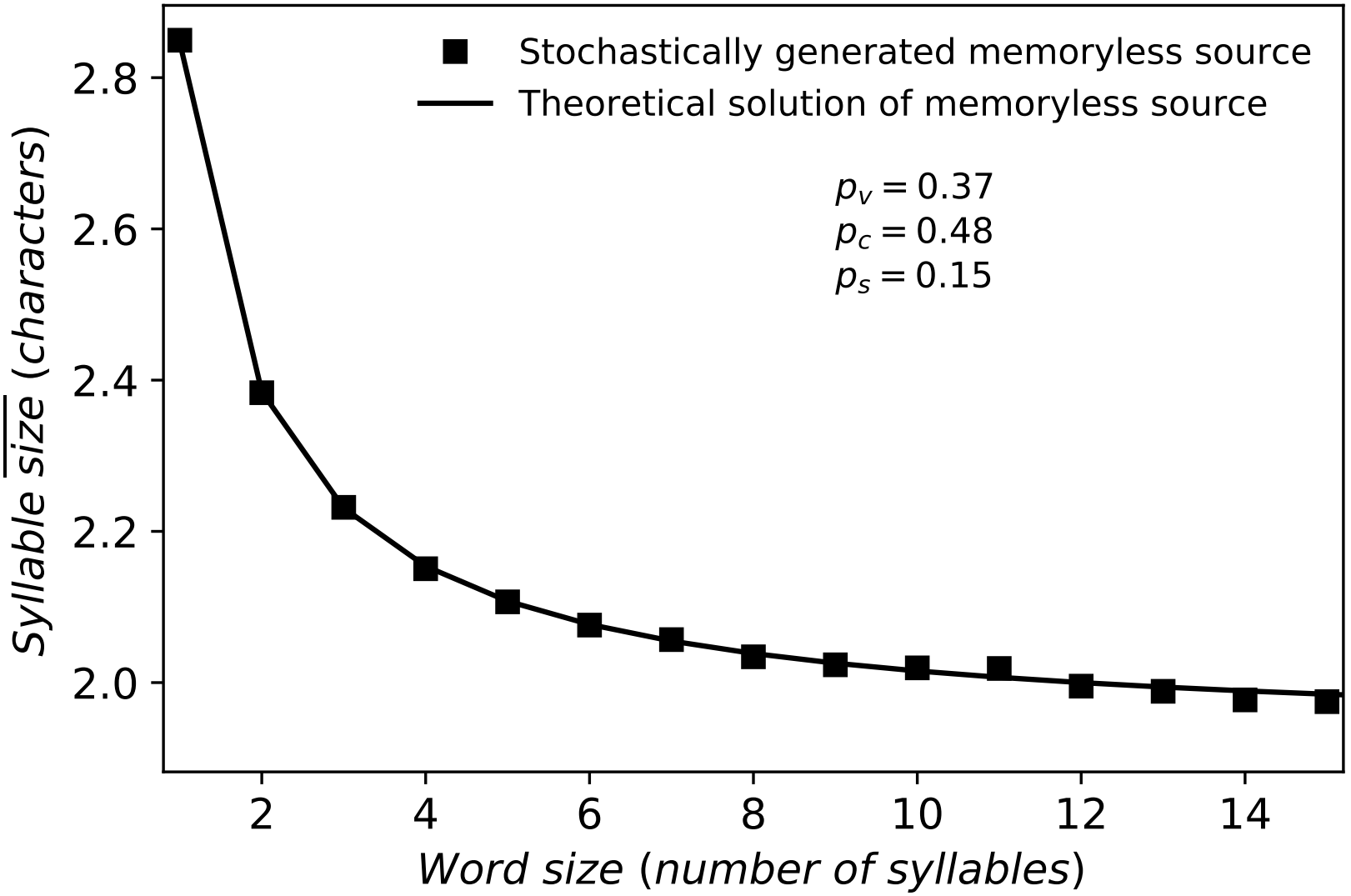
$$= (m+1)\sum_{n=0}^{\infty} nP(N_0 = n)$$

$$= (m+1)\sum_{n=0}^{\infty} np_c^n(1-p_c) = \frac{p_c}{1-p_c}(m+1),$$

As a result, we obtain **Menzerath's law**:

$$\mathbb{E}\left(\frac{N+M}{M}\middle|M = m\right) = \frac{a}{m} + b, \quad a = \frac{p_c}{1-p_c}, \quad b = 1 + \frac{p_c}{1-p_c}.$$

The plot for monkey typing



Standardized Gutenberg Corpus

We have used the Standardized Gutenberg Corpus database:

- a curated open access version of Project Gutenberg,
- over 50 000 books in diverse language.

We have selected:

- 21 languages that use the Latin alphabet and are represented by at least 5 books,
- up to 2500 randomly selected books per language.

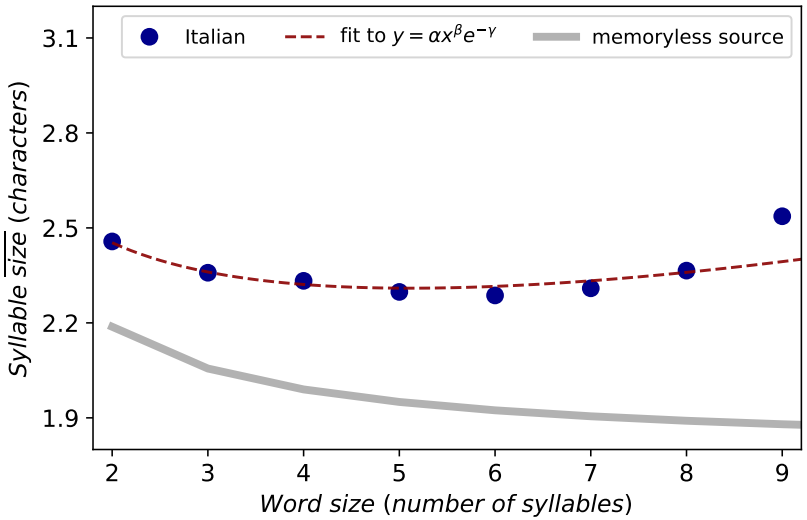
Syllabification — a language-independent procedure

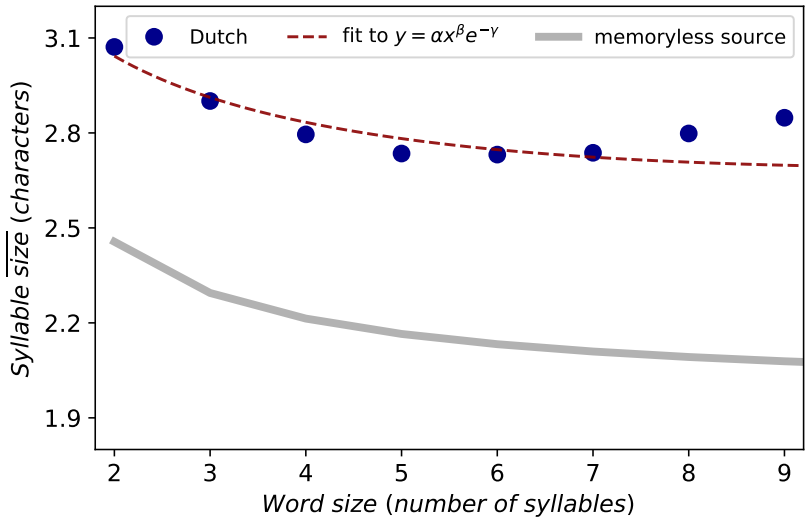
Sonority Sequencing Principle:
The nucleus is the element of the maximum sonority, the onset and the coda have a decreasing sonority. The sonority scale:

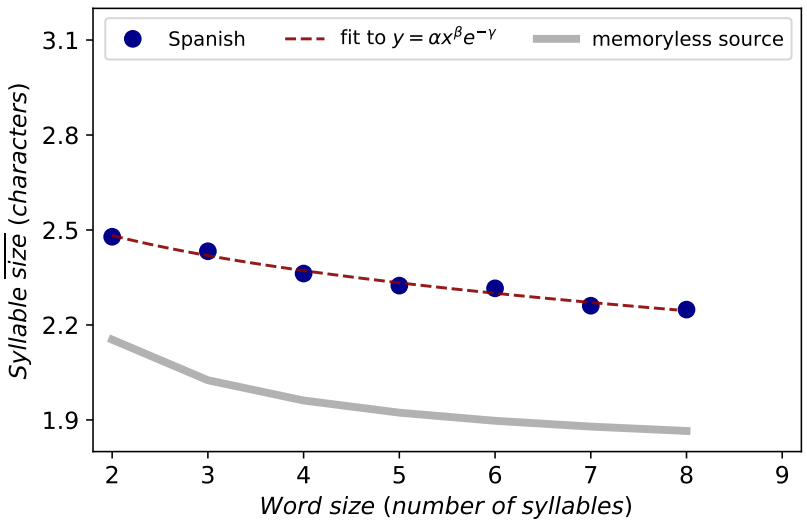
vowel > approximant > liquid > nasal > fricative > affricate > occlusive

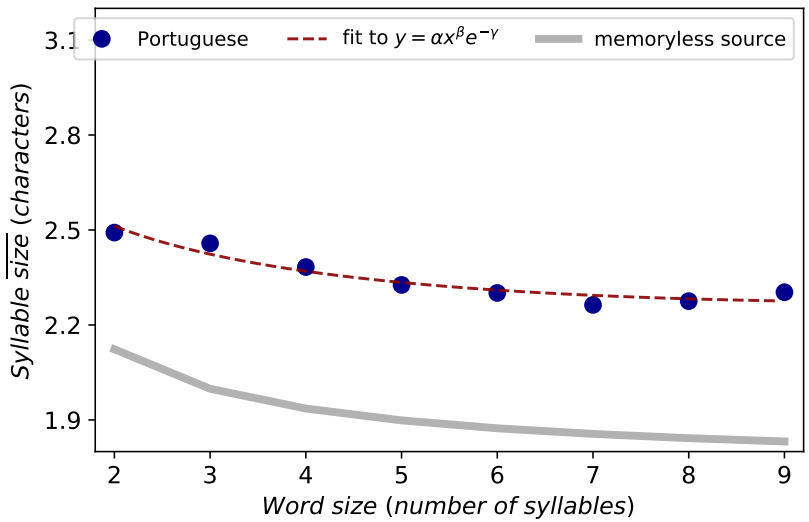
Sonority class	Graphemes
Vowels	a, e, i, o, u, y, à, á, â, ä, æ, ã, å, ā, ą, è, é, ê, ë, ē, ė, ę, î, ï, í, ī, į, ì, ô, ö, ò, ó, œ, ø, ō, õ, û, ü, ù, ú, ū, ů, ÿ, ű, ő, ŵ, ŷ, ě, ý, ǫ
Approximants	ŭ, w, ł
Liquids	l, r, ř
Nasals	m, n, ñ, ń, ŋ, ň,
Fricatives	ß, z, v, s, f, ç, ć, ś, ŝ, ĉ, ĥ, h, ĵ, š, ž, ð, đ
Affricates	x, j, ź, ż, ĝ, č
Occlusives	b, c, d, g, t, k, p, q, þ, ď, ť

Real syllable length vs. monkey-typing prediction





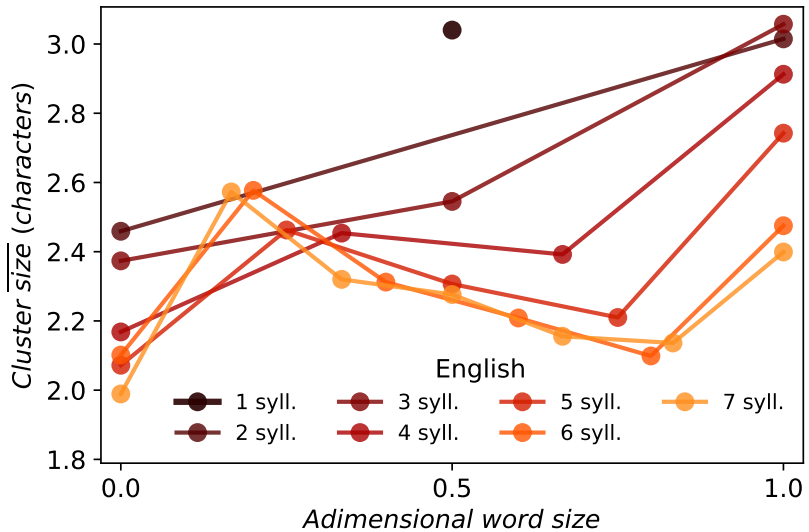


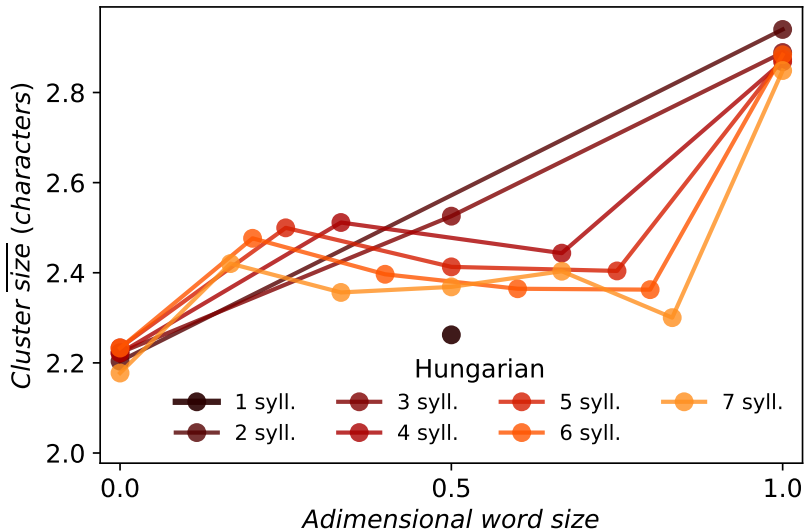


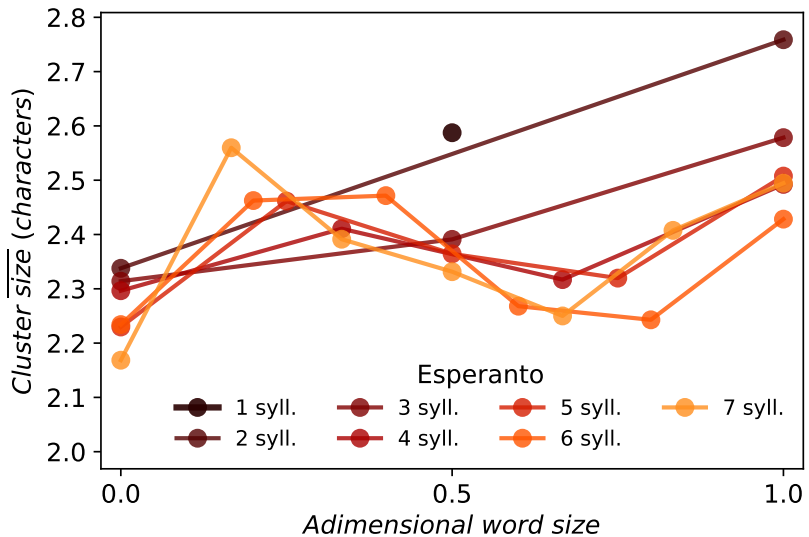
The inverted regime in Menzerath-Altmann's law!

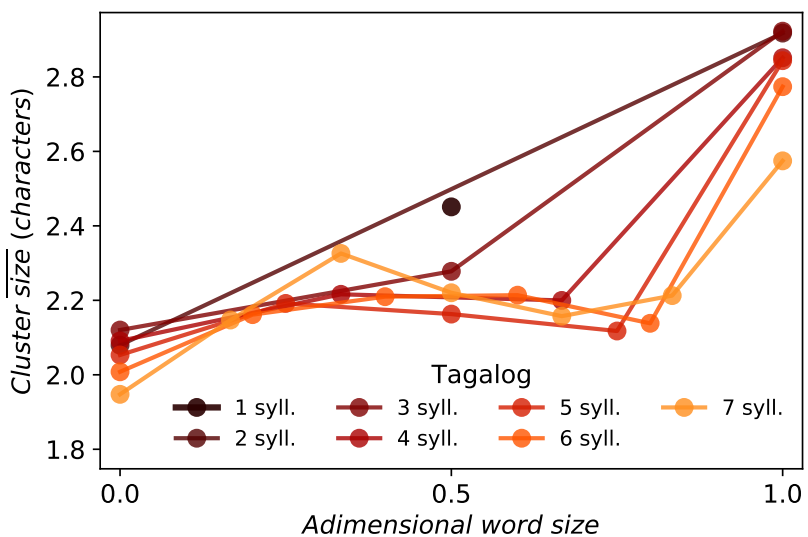
Language	α	β	γ	R^2	β/γ
English	3.19	$-3.5 \cdot 10^{-1}$	$-5.4 \cdot 10^{-2}$	0.82	6.5
French	2.96	$-2.4 \cdot 10^{-2}$	$2.6 \cdot 10^{-2}$	0.94	-
Finnish	2.69	$-5.6 \cdot 10^{-2}$	$-8.8 \cdot 10^{-3}$	0.63	6.4
German	3.19	$-1.5 \cdot 10^{-1}$	$-2.4 \cdot 10^{-2}$	0.62	6.3
Italian	2.59	$-1.8 \cdot 10^{-1}$	$-3.6 \cdot 10^{-2}$	0.41	5.1
Dutch	3.27	$-1.4 \cdot 10^{-1}$	$-1.3 \cdot 10^{-2}$	0.63	10.9
Spanish	2.6	$-5.3 \cdot 10^{-2}$	$4.6 \cdot 10^{-3}$	0.98	-
Portuguese	2.66	$-1.1 \cdot 10^{-1}$	$-1.0 \cdot 10^{-2}$	0.93	10.9
Hungarian	2.69	$-8.9 \cdot 10^{-2}$	$-1.1 \cdot 10^{-2}$	0.81	7.9
Swedish	2.82	$-1.0 \cdot 10^{-1}$	$-2.0 \cdot 10^{-2}$	0.71	5.1
Esperanto	2.69	$-1.4 \cdot 10^{-1}$	$-2.0 \cdot 10^{-2}$	0.95	7.1
Latin	2.82	$-1.7 \cdot 10^{-1}$	$-2.1 \cdot 10^{-2}$	0.95	8.2
Danish	2.77	$-5.2 \cdot 10^{-2}$	$-6.5 \cdot 10^{-3}$	0.44	8.1
Tagalog	2.69	$-1.1 \cdot 10^{-1}$	$-3.9 \cdot 10^{-3}$	0.98	28.7
Catalan	2.83	$-1.9 \cdot 10^{-1}$	$-2.5 \cdot 10^{-2}$	0.96	7.5
Polish	3.13	$-1.6 \cdot 10^{-1}$	$-4.1 \cdot 10^{-3}$	0.99	39.5
Norwegian	2.87	$-1.4 \cdot 10^{-1}$	$-2.7 \cdot 10^{-2}$	0.67	5.2
Czech	2.76	$-2.3 \cdot 10^{-1}$	$-3.5 \cdot 10^{-2}$	0.95	6.7
Welsh	3.3	$-2.8 \cdot 10^{-3}$	$5.9 \cdot 10^{-2}$	0.99	-
Icelandic	2.67	$5.8 \cdot 10^{-2}$	$1.9 \cdot 10^{-2}$	0.13	3.0
Afrikaans	3.4	$-3.0 \cdot 10^{-1}$	$-4.8 \cdot 10^{-2}$	0.97	6.2

Cluster length vs. position in the word









Conclusion

- We have shown that a monkey-typing source is able to reproduce Menzerath's law.
- This observation questions extant interpretations of Menzerath's law for communication complexity and efficiency.
- However, we have also shown that natural languages exhibit an inverted regime in Menzerath-Altmann's law (MAL).
- Thus the more specific MAL can be a property that distinguishes human languages from monkey typing.

G. Torre I., Dębowski Ł., Hernández-Fernández A. (2021) Can Menzerath's law be a criterion of complexity in communication? PLoS ONE 16(8): e0256133