sciendo
DOI 10.2478/jazcas-2025-0032

LONGER WORDS, EASIER-TO-PRONOUNCE PHONEMES: A PILOT STUDY

JÁN MAČUTEK[1] – RADEK ČECH[2] – MICHAELA KOŠČOVÁ[3]
[1]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
(ORCID: 0000-0003-1712-4395)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-4412-4588)
[3]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia
(ORCID: 0000-0002-5541-1224)

MAČUTEK, Ján – ČECH, Radek – KOŠČOVÁ, Michaela: Longer Words, Easier-to-pronounce Phonemes: A Pilot Study. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 355 – 365.

Abstract: The study investigates the relationship between word length and phoneme sonority in six languages across diverse language families. Building on the principle of least effort and the Menzerath-Altmann law, the research is aimed to analyze the phoneme sonority using translated New Testament texts in Bilua, Bola, Czech, Gagauz, Jamamadi, and Tongan. The findings reveal that in languages with complex syllables, the tendency of longer words to contain shorter syllables—consistent with the Menzerath-Altmann law—results in a higher proportion of vowels, thereby increasing the mean phoneme sonority. In contrast, languages with simple syllable structures exhibit either a decrease in mean phoneme sonority or no clear trend. Further, mean consonant sonority increases with word length in Bilua, Czech, and Gagauz, while no clear trend is observed in Bola, Jamamadi, and Tongan. Conversely, mean vowel sonority increases with word length in Bola, Jamamadi, and Tongan, but remains stable or decreases in Bilua, Czech, and Gagauz. Overall, the analysis reveals consistent patterns linking word length and sonority across all six languages.
Keywords: principle of least effort, phoneme, sonority, lenition, word length, syllable

1 INTRODUCTION

The principle of least effort (Zipf 1949) is one of key forces shaping structure and properties of language units. According to the principle, in language there is a tendency towards economizing that is probably the most easily noticeable in the relationship between frequency and length of language units. The well-known Zipf's law of abbreviation is usually formulated for words (the higher the frequency of a word, the shorter it tends to be, see Ferrer-i-Cancho et al. 2022), but the same holds true e.g. also for syllables (Rujević et al. 2021) and lengths of dependency distances (Chen and Gerdes 2022).

However, the validity of the principle of least effort is not restricted only to frequencies. The Menzerath-Altmann law (Menzerath 1954; Altmann 1980) states

that longer units are composed of parts that are on average shorter (e.g. longer words consist of shorter syllables). It is another manifestation of the principle of least effort – if we must use longer words, we build them from simpler syllables. But the law refers to types (see Motalová et al. 2023 and Wang and Kelih 2024 for reasons why it is not valid for tokens), and thus disregards frequencies.

Similarly, the tendency of syllables to shorten is not the only possible way how to reduce effort in longer words. Already Hřebíček and Altmann (1996, p. 55) wrote that one could use "less complicated" instead of "shorter" parts in the formulation of the Menzerath-Altmann law. In this paper, we present some tendencies in the relationship between word length and phoneme sonority in six languages. Longer words tend to contain phonemes that are easier to pronounce (either in absolute terms, or relatively with respect to their neighbours).

The study is motivated by the fact that some languages allowing only simple syllable structure (i.e. only CV and V syllables, see Maddieson 2007, p. 96) display non-standard behaviour with respect to the Menzerath-Altmann law, see Mačutek et al. (2025). The tendency to use easier-to-pronounce phonemes in longer words seems to be universal regardless of syllable types allowed in particular languages.

We emphasize that we analyze words on the phonological, and not on the phonetic level, e.g. we consider theoretical properties of phonemes in written texts, and not physical properties of sounds in actual utterances.

2 METHODOLOGY AND LANGUAGE MATERIAL

2.1 Sonority hierarchy

Phonemes in particular languages are ranked according to sonority hierarchy, see Tab. 1. We follow mostly Szigetvári (2008, p. 96); in addition, fricatives are merged into one category with plosives (Parker 2011 writes that "...the placement of affricates between stops and fricatives is a controversial issue, remaining open to disagreement. Many scales either leave affricates out entirely or group them with plosives…"). In diphthongs, both vowels are taken into account.

phonemes	sonority index
low vowels	10
mid vowels	9
high vowels and semivowels	8
rhotics	7
laterals	6
nasals	5
voiced fricatives	4
voiceless fricatives	3
voiced plosives and affricates	2
voiceless plosives and affricates	1

Tab. 1. Sonority hierarchy

Any sonority hierarchy provides only a ranking of phonemes. The ranks do not reflect actual differences (e.g. the difference between voiceless fricatives and voiced fricatives does not have to be the same as the one between rhotics and high vowels). Anyway, it can be used to characterize the mean sonority of phonemes in words.

2.2 Language material

As language material, we use translations of the New Testament (27 books) into six languages from five different language families: Bilua (from the Central Solomon language family), Bola (Austronesian), Czech (Indo-European), Gagauz (Turkic), Jamamadi (Arawan), and Tongan (Austronesian). The Bible as a source of texts has its drawbacks (e.g. there are many proper names especially of Greek, Hebrew, and Latin origin), but for many languages it is the only easily available collection of texts that are long enough to enable statistical analyses. Book titles, references to other sources etc. were deleted. Links to Bible translations can be found in Tab. 2 (if the webpage provides access only to individual chapters, the link to the first chapter of the Gospel of Matthew is given).

language	link
Bilua	https://www.bible.com/bible/2979/MAT.1.BLBNT
Bola	https://www.scriptureearth.org/data/bnp/PDF/00-PBIbnp-web.pdf
Czech	https://bible.jecool.net/wp-content/uploads/2016/03/bible-velka.pdf
Gagauz	https://www.bible.com/en-GB/bible/2554/MAT.1.GAGNTL
Jamamadi	https://www.bible.com/bible/3158/MAT.1.JAANT
Tongan	https://ebible.org/pdf/ton/ton_nt.pdf

Tab. 2. Links to texts used

Four of these languages have only simple syllables. In Bilua (Obata 2003), all monosyllables are of the CV structure. In longer words, the first syllable can be CV or V, with all other syllables being CV. Bola (van den Berg and Wiebe 2019), Jamamadi (Dixon and Vogel 2004), and Tongan (Garellek and Tabain 2020) have only CV and V syllables without positional restrictions. Words containing syllables of other types (e.g. toponyms like *Nasaret* ‘Nazareth’ in Bola) were removed.

On the other hand, Czech (Short 1993) and Gagauz (Pokrovskaja 1964) allow also more complex syllables (and, consequently, consonant clusters exist in these two languages).

As we focus on the phonological analyses of written texts, it is important to note that all these languages have shallow orthographies, i.e. the phoneme-grapheme ratios are close to one-to-one (see Coulmas 2002, pp. 101–102). Therefore, the phonological transcriptions are relatively easy to do.

3 RESULTS

In order to guarantee certain stability of the means, in the following tables and figures only those word lengths are presented for which at least ten different words occur in the text.

3.1 Menzerath-Altmann law

We first present results of the analysis of the Menzerath-Aktmann law (see Tab. 3 and Fig. 1), as they are needed to understand the development of the mean sonority in the next sections. Data for Bilua, Bola, Jamamadi, and Tongan are taken from Mačutek et al. (2025). We add the results for Czech and Gagauz. The mean syllable length decreases with the increasing word length in languages with complex syllables (Czech, Gagauz). Languages with only simple syllables (Bilua, Bola, Jamamadi, Tongan) do not display a clear Menzerathian tendency.

word length in syllables	mean syllable length in phonemes					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	1.96	1.90	3.52	2.88	1.92	1.88
2	1.94	1.93	2.70	2.51	1.91	1.87
3	1.96	1.88	2.41	2.41	1.93	1.86
4	1.96	1.89	2.24	2.38	1.94	1.84
5	1.96	1.90	2.17	2.35	1.96	1.84
6	1.97	1.86	2.18	2.31	1.96	1.86
7	1.96		2.14	2.26	1.97	1.86
8					1.97	1.83
9					1.97	1.82
10					1.97	
11					1.97	

Tab. 3. Relationship between word length and the mean syllable length

3.2 Mean phoneme sonority

The mean phoneme sonority (Tab. 4, Fig. 2) decreases with the increasing word length in Bilua. It is a consequence of the syllable structure in this language. As only the first syllable can be V and all other must be CV, the proportion of consonants increases, and, as consonants are less sonorous then vowels, the mean sonority decreases.

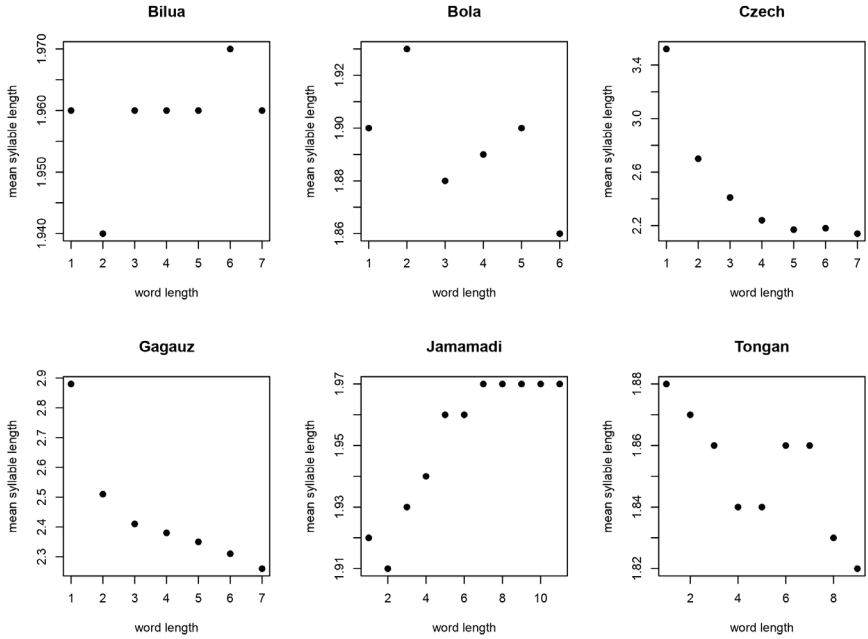

Fig. 1. Relationship between word length and the mean syllable length

In Czech and Gagauz, we observe the opposite trend, the mean phoneme sonority increases with the increasing word length. It is a consequence of the Menzerath-Altmann law – syllables get shorter, which means a higher proportion of vowels, and vowels have a higher sonority than consonants.

No clear trend is visible in Bola, Jamamadi, and Tongan.

word length in syllables	mean phoneme sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	6.88	6.47	5.22	5.45	6.67	6.19
2	6.61	6.42	5.68	5.89	6.55	6.33
3	6.60	6.66	5.92	6.14	6.48	6.28
4	6.47	6.54	6.14	6.26	6.53	6.27
5	6.44	6.52	6.26	6.37	6.57	6.27
6	6.47	6.63	6.13	6.40	6.59	6.19
7	6.48		6.20	6.44	6.62	6.21
8					6.62	6.25
9					6.64	6.31
10					6.59	
11					6.47	

Tab. 4. Relationship between word length and the mean phoneme sonority

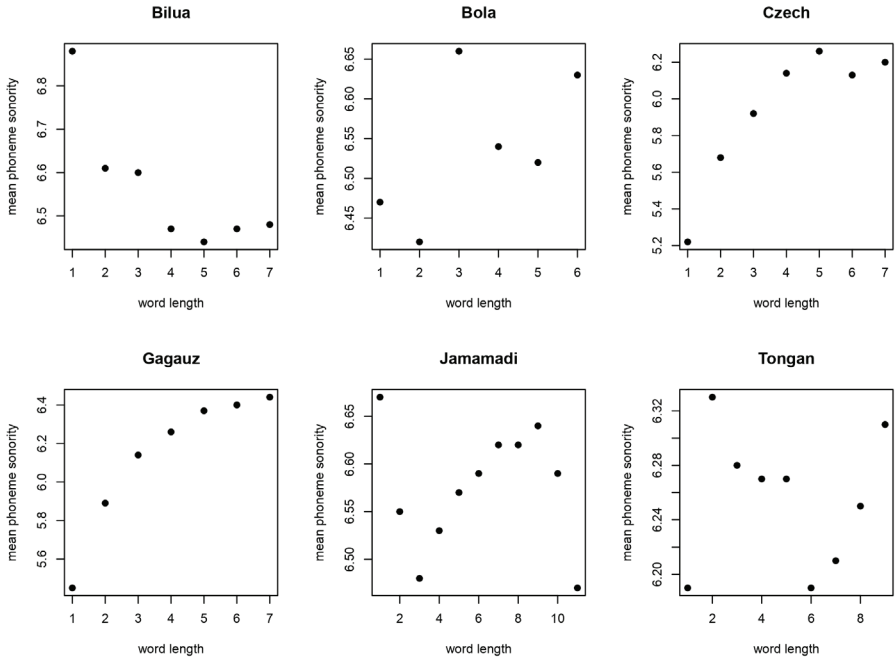

Fig. 2. Relationship between word length and the mean phoneme sonority

3.3 Mean consonant sonority

The mean consonant sonority increases in Bilua, Czech, and Gagauz (see Tab. 5 and Fig. 3). An analogy with lenition offers itself to explain this observation.

In Bilua, all consonants, possibly except for the one at the beginning of a word, are in intervocalic positions. As the mean syllable length decreases when words get longer, the probability that a consonant is in an intervocalic position in Czech and Gagauz increases.

According to Kirchner (2001, p. 138), "[i]t is fairly well established that intervocalic position is a natural lenition environment", and lenition closely correlate with increasing sonority (i.e. voicing is one of exemplifications of lenition). While we cannot apply this term literally (we observe neither a diachronic development of a language, nor preferring lenited consonants by individual speakers), we can say that Bilua, Czech, and Gagauz prefer more sonorous consonants in intervocalic positions, and these positions occur more often in longer words. These findings are in line with the paper by File-Muriel (2016) who reports an increasing lenition rates of /s/ in a Colombian variety of Spanish (though there is also an important difference – he works with tokens, not with types).

The mean consonant sonority behaves quite chaotically, or at least with a much less clear trend, in Bola, Jamamadi, and Tongan. All consonants that are not at the beginning of words are in intervocalic positions too (as in Bilua), but these languages allow also vocalic clusters (which Bilua forbids).

word length in syllables	mean consonant sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	3.22	3.46	3.63	3.08	3.79	2.67
2	3.34	3.63	3.73	3.55	3.66	3.12
3	3.45	3.83	3.77	3.86	3.51	3.02
4	3.40	3.61	3.94	4.06	3.67	2.89
5	3.45	3.48	3.99	4.17	3.79	2.85
6	3.54	3.51	3.86	4.19	3.86	2.71
7	3.53		3.99	4.27	3.89	2.66
8					3.90	2.68
9					3.93	2.71
10					3.80	
11					3.60	

Tab. 5. Relationship between word length and the mean consonant sonority

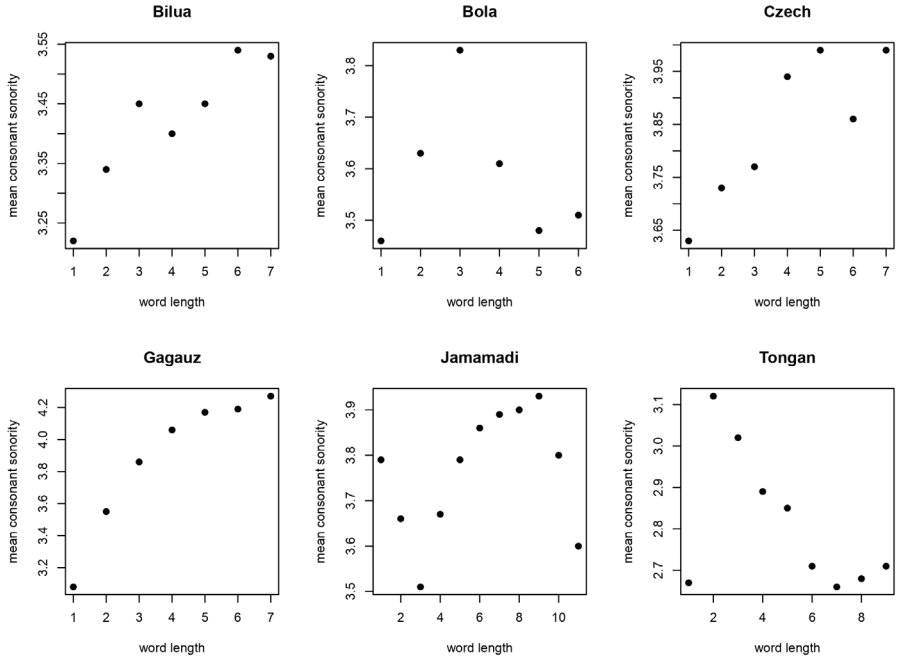

Fig. 3. Relationship between word length and the mean consonant sonority

3.4 Mean vowel sonority

The exact opposite of Section 3.3 is mean vowel sonority (see Tab. 6 and Fig. 4) – the mean sonority of vowels increases with the increasing word length in Bola, Jamamadi (admittedly, not so regularly), and Tongan, i.e. in languages in which there was no trend in the mean consonant sonority. On the other hand, Bilua, Czech, and Gagauz, languages with a systematic increase in consonant sonority, the mean vowel sonority does not increase (it rather decreases in Bilua and Gagauz, and behaves irregularly in Czech).

word length in syllables	mean vowel sonority					
	Bilua	Bola	Czech	Gagauz	Jamamadi	Tongan
1	8.97	8.92	8.86	8.95	9.19	9.00
2	9.02	8.94	8.82	8.90	9.12	9.00
3	8.98	9.08	8.85	8.88	9.21	9.01
4	8.97	9.12	8.86	8.89	9.22	9.05
5	8.96	9.21	8.90	8.90	9.21	9.09
6	8.91	9.30	8.83	8.89	9.22	9.15
7	8.94		8.83	8.86	9.26	9.22
8					9.25	9.22
9					9.28	9.24
10					9.28	
11					9.23	

Tab. 6. Relationship between word length and the mean vowel sonority

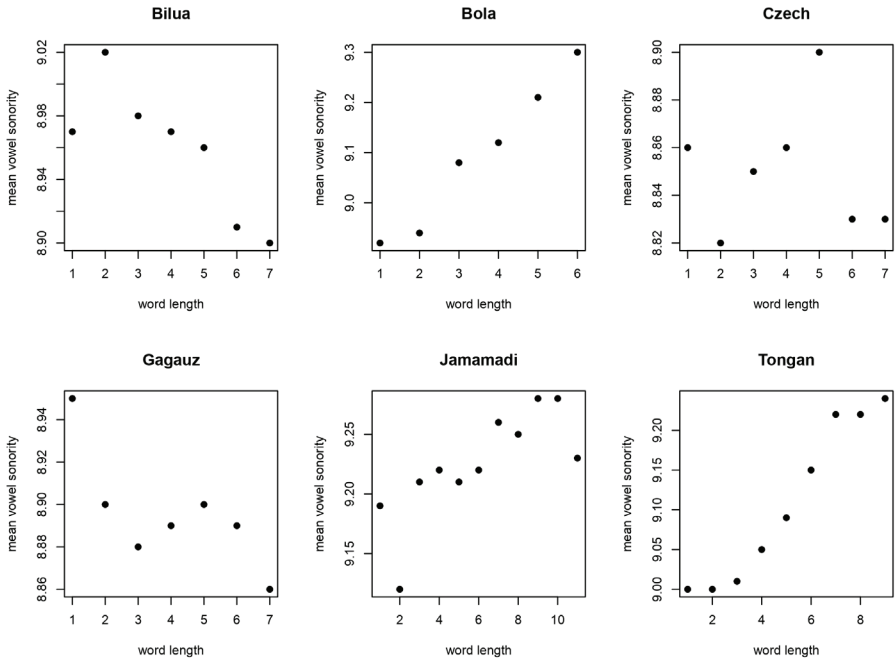

Fig. 4. Relationship between word length and the mean vowel sonority

4 CONCLUSION AND DISCUSSION

There is a systematic relationship between word length and sonority in all six languages under analysis. The nature of this relationship seems to depend on the syllable types allowed in individual languages. However, there is a connection to the principle of least effort in all six cases.

Czech and Gagauz have complex syllables and consonant clusters occur in them. As a consequence of the Menzerath-Altmann law, syllables are shorter in longer words. This means that consonant clusters are less probable in longer words, and syllalbes (and phonemes in them) become less difficult to be pronounced (see e.g. Stoel-Gammon 2010, p. 273). The Menzerath-Altmann law explains also why the mean phoneme sonority increases as words get longer – shorter syllables have higher proportions of vowels, and vowels are more sonorous then consonants.

In Bilua, due to the positional restrictions of its syllables types (V only at the beginning of words that have at least two syllables, CV elsewhere), the proportion of consonants increases in longer words. Therefore, there is a negative correlation between word length and the mean sonority of phonemes.

We can observe a positive correlation between word length and the mean sonority of consonants in Bilua, Czech, and Gagauz. According to Gurevich (2011), "[v]oicing, for example, has an explanation rooted in the laws of physics, specifically aerodynamics: intervocalically the vocal cords may continue to vibrate after the first vowel, through the consonant, and into the second vowel". Voicing is a typical exemplification of lenition, and voiced consonants are higher in the sonority hierarchy then their unvoiced counterparts. The increasing sonority of consonants in intervocalic positions thus weakens articulatory effort and thus compensates for increasing word length. And indeed, consonants in intervocalic positions are more probable in longer words – in Czech and Gagauz as a consequence of the Menzerath-Altmann law, and in Bilua because of more intervocalic slots available for consonants.

In Bola and Tongan (and to a slightly lesser extent also in Jamamadi) there is a positive correlation between word length and the mean vowel sonority. It means that lower vowels are preferred in longer words. But according to Jaeger (1978, p. 313), "the narrower constriction for high vowels causes air pressure in the oral cavity to be greater than that during low vowel", and (Napoli et al. 2014, p. 427) "[c]onsequently, more pulmonic effort is needed for the airflow across the glottis to overcome the resistive force of the oral cavity's higher air pressure". Thus, low vowels (with a higher sonority) are preferred because they require less effort.

Many questions appear with every (incomplete) answer. The impact of phoneme inventory size and structure must be investigated (e.g. if a language has more pairs of voiced and unvoiced consonants, it can make utterances easier by voicing consonants in intervocalic positions; if not, it can prefer lowering of vowels). We

focused here on types – tokes must be studied too. And more languages must be analysed before one can reach a conclusion. But this study confirms once more that the least effort principle is (almost) ubiquitous in language.

The principle of least effort is, however, a double-edged sword. For a speaker without a hearer, it would be the most comfortable to utter only easy-to-pronounce phonemes (e.g. only low vowels). But e.g. CV syllables with a higher difference in sonority of a consonant and a vowel are easier to segment for a hearer (Yavas and Gogate 1999). Being too "lazy", a speaker would risk information loss at a hearer's side and a necessity of re-sending a message, which would require another effort. Thus, language finds itself in a state of a Zipfian equilibrium (Zipf 1935) in which the speaker's drive to economize is controlled by the hearer's feedback.

ACKNOWLEDGEMENTS

Supported by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00748 (J. Mačutek), by the European Regional Development Fund project "A lifetime with language: the nature and ontogeny of linguistic communication (LangInLife)" (reg. no.: CZ.02.01.01/00/23_025/0008726) (R. Čech), and by projects APVV-21-0216 and VEGA 2/0120/24 (M. Koščová).

References

Altmann, G. (1980). Prolegomena to Menzerath's law. In: R. Grotjahn (ed.): Glottometrika 2. Bochum: Brockmeyer, pp. 1–10.

Chen, X., and Gerdes, K. (2022). Dependency distances and their frequencies in Indo-European language. Journal of Quantitative Linguistics, 29(1), pp. 106–125.

Coulmas, F. (2002). Writing systems. An introduction to their linguistic analysis. Cambridge: Cambridge University Press. 270 p.

Dixon, R. M. W., and Vogel, A. R. (2004). The Jarawara language of Southern Amazonia. Oxford: Oxford University Press, 636 p.

Ferrer-i-Cancho, R., Bentz, C., and Seguin, C. (2022). Optimal coding and the origin of Zipfian laws. Journal of Quantitative Linguistics, 29(2), pp. 165–194.

File-Muriel, R. J. (2010). Lexical frequency as a scalar variable in explaining variation. The Canadian Journal of Linguistics, 55(1), pp. 1–25.

Garellek, M., and Tabain, M. (2020). Tongan. Journal of the International Phonetic Association, 50(3), pp. 406–413.

Gurevich, N. (2011). Lenition. In: M. van Oostendorp – C. J. Ewen – E. Hume – K. Rice (eds.): The Blackwell companion to phonology. Chichester: Wiley – Blackwell, pp. 1559–1575.

Hřebíček, L., and Altmann, G. (1996). Levels of order in language. In: P. Schmidt (ed.): Glottometrika 15. Issues in general linguistic theory and the theory of word length. Trier: WVT, pp. 38–61.

Jaeger, J. J. (1978). Speech aerodynamics and phonological universals. In: J. J. Jaeger – A. C. Woodbury – F. Ackerman – C. Chiarello – O. D. Gensler – J. Kingston – E. E. Sweetser – H. Thompson – K. W. Whistler (eds.): Proceedings of the 4th annual meeting of the Berkeley Linguistics Society. Berkeley (CA): Berkeley Linguistics Society, pp. 312–329.

Kirchner, R. (2001). An effort based approach to consonant lenition. Abingdon: Routle.g. 303 p.

Mačutek, J., Nogolová, M., Rovenchak, A., and Čech, R. (2025). What does the Menzerath-Altmann law really say? Journal of Quantitative Linguistics (accepted paper).

Maddieson, I. (2007). Issues in phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories, and tone contrasts. In: M.-J. Solé – P. Beddor Speeter – M. Ohala (eds.): Experimental approaches to phonology. Oxford: Oxford University Press, pp. 93–103.

Menzerath, P. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler. 131 p.

Motalová, T., Mačutek, J., and Čech, R. (2023). Word length in Chinese: The Menzerath-Altmann law is valid after all. Journal of Quantitative Linguistics, 30(3–4), pp. 301–321.

Napoli, D. J., Sanders, N., and Wright, R. (2014). On the linguistic effects of articulatory ease, with a focus on sign languages. Language, 90(2), pp. 424–456.

Obata, K. (2003). A grammar of Bilua: A Papuan language of the Solomon Islands. Canberra: The Australian National University. 333 p.

Parker, S. (2011). Sonority. In: M. van Oostendorp – C. J. Ewen – E. Hume – K. Rice (eds.): The Blackwell companion to phonology. Chichester: Wiley – Blackwell, pp. 1160–1184.

Pokrovskaja, L. A. (1964). Grammatika gagauzskogo jazyka (A grammar of the Gagauz language). Moskva: Nauka. 300 p.

Rujević, B., Kaplar, M., Kaplar, S., Stanković, R., Obradović, I., and Mačutek, J. (2021). Quantitative analysis of syllables in Croatian, Serbian, Russian, and Ukrainian. In: A. Pawłowski – J. Mačutek – S. Embleton – G. Mikros (eds.): Language and text: Data, Models, Information and Applications. Amsterdam, Philadelphia: Benjamins, pp. 55–67.

Short, D. (1993). Czech. In: B. Comrie – G. G. Corbett (eds.): The Slavonic languages. London: Routle.g. pp. 455–532.

Stoel-Gammon, C. (2010). The word complexity measure: Description and application to developmental phonology and disorders. Clinical Linguistics & Phonetics, 24(4–5), pp. 271–282.

Szigetvári, P. (2008). What and when? In: J. Brandão de Carvalho – T. Scheer – P. Ségéral (eds.): Lenition and fortition. Berlin, New York: de Gruyter, pp. 93–129.

van den Berg, R., and Wiebe, B. (2019). Bola grammar sketch. Ukarumpa: SIL-PNG Academic Publications, 292 p.

Wang, T., and Kelih, E. (2024). Boundary conditions for the Menzerath-Altmann law. What should be taken: Tokens, types or lemmas? Glottometrics 57, pp. 1–20.

Yavas, M., and Gogate, L. J. (1999). Phoneme awareness in children: A function of sonority. Journal of Psycholinguistic Research, 28(3), pp. 245–260.

Zipf, G. K. (1949). Human behavior and the principle of least effort. Cambridge (MA): Addison-Wesley Press, 573 p.

Zipf, G. K. (1935). The psycho-biology of language. Boston: Houghton-Mifflin, 336 p.