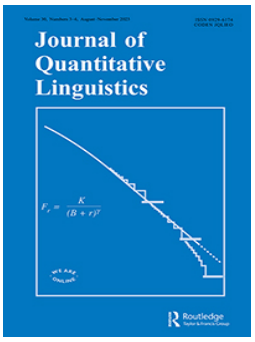

Journal of Quantitative Linguistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/njql20

Word Length in Chinese: The Menzerath-Altmann Law is Valid After All

Tereza Motalová, Ján Mačutek & Radek Čech

To cite this article: Tereza Motalová, Ján Mačutek & Radek Čech (2023) Word Length in Chinese: The Menzerath-Altmann Law is Valid After All, Journal of Quantitative Linguistics, 30:3-4, 304-321, DOI: 10.1080/09296174.2023.2259937

To link to this article: https://doi.org/10.1080/09296174.2023.2259937

Published online: 06 Nov 2023.

Submit your article to this journal

Article views: 147

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=njql20

JOURNAL OF QUANTITATIVE LINGUISTICS
2023, VOL. 30, NOS. 3–4, 304–321
https://doi.org/10.1080/09296174.2023.2259937

Routledge
Taylor & Francis Group

Check for updates

Word Length in Chinese: The Menzerath-Altmann Law is Valid After All

Tereza Motalová[a], Ján Mačutek[b,c] and Radek Čech[d]

[a]Department of Asian Studies, Faculty of Arts, Palacky University Olomouc, Olomouc, Czech Republic; [b]Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia; [c]Department of Mathematics, Faculty of Natural Sciences and Informatics, Constantine the Philosopher University in Nitra, Nitra, Slovakia; [d]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic

ABSTRACT
According to the Menzerath-Altmann law, longer language constructs consist, on average, of shorter constituents. It is most often studied at the level of words and syllables (the mean syllable length gets shorter with the increasing word length). Its validity at this level was corroborated in several languages. However, it was claimed that Chinese is an exception with respect to the validity of the Menzerath-Altmann law. We show that the law is valid if word types are considered, while the behaviour of word tokens is different. This difference can be explained by the fact that the Zipf law of abbreviation is valid not only for words but also for syllables (shorter syllables are used more frequently).

KEYWORDS word length; Menzerath-Altmann law; Chinese; syllable; Chinese characters

1 Introduction

'The longer a language construct, the shorter its components (constituents)' (Altmann, 1980, p. 124). This statement, today known as the Menzerath-Altmann law (MAL henceforward), describes how lengths of constructs and constituents (which are understood as the immediate lower neighbours of constructs in the hierarchy of language units) influence each other. The usual mathematical formulation of the MAL is

$$y(x) = ax^b \tag{1}$$

where x is the length of the construct, $y(x)$ is the mean length of constituents in constructs of length x, and a, b are parameters.[1] The goodness-of-fit is usually expressed in terms of the determination coefficient R^2 (see Mačutek & Wimmer, 2013). A typical example is the relation between lengths of words and syllables the words consist of, i.e. the more syllables a word contains, the shorter on average its syllables are (measured in phonemes). The law can be

CONTACT Tereza Motalová tereza.motalova@upol.cz

© 2023 Informa UK Limited, trading as Taylor & Francis Group

traced back at least to Menzerath (1954). It appeared in a slightly different context already in Menzerath and de Oleza (1928).

Over the course of the last four decades, since Altmann (1980) generalized the observation by Menzerath (1954), the MAL has enjoyed an ever-increasing attention from scientists. Studies of different language units in many languages have been conducted. An overview of older results can be found in Cramer (2005). From the more recent ones, we mention the relation between lengths of words and morphemes (Pelegrinová et al., 2021; Stave et al., 2021), canonical word forms and syllables (Mačutek & Rovenchak, 2011), word length motifs and words (Mačutek & Mikros, 2015), clauses and phrases (Mačutek et al., 2017) and sentences and clauses Kulacka (2009); Wang and Čech (2016); Jin and Liu (2017).[2]

Although the scope of the MAL was substantially broadened, the relation between word length and the mean syllable length remains its most often studied exemplification. However, one must distinguish between word types and tokens (i.e. occurrences of word types) when studying this relation. Menzerath (1954) observed the tendency of syllables to shorten with the increasing word length in language material from a German dictionary (where it does not make any sense to consider tokens). The validity of the law for word types was later confirmed in several languages: e.g. Czech (Kelih, 2008; Milička, 2014), English (Cramer, 2005), Greek (Mikros & Milička, 2014), Italian (Cramer, 2005), Indonesian (Cramer, 2005), Macedonian (Kelih, 2008), Maninka (Rovenchak, 2015[3]), Romanian (Dinu & Dinu, 2009), Russian (Kelih, 2008), Serbian/Serbo-Croatian/Croatian (Cramer, 2005; Grzybek, 1999; Kelih, 2010), Slovene (Kelih, 2008, 2012) and Ukrainian (Buk and Rovenchak 2007).

But several works reveal a different pattern of data obtained from word tokens. Thus, according to Mikros and Milička (2014), Milička (2014) and Rovenchak (2015), the mean syllable length computed from word tokens does not abide by the MAL, although syllables in word types from the same texts follow it. Mačutek et al. (2019) investigate orthographic transcriptions of spoken Czech and report that the MAL fits the data well in the majority of texts, but its validity is not general. Rujević et al. (2021) use tokens in four languages (Croatian, Serbian, Russian, Ukrainian) and show that the decreasing tendency of the mean syllable length cannot be observed in the data. Kraviarová and Zimmermann (2010) and Andres et al. (2012) seem not to be aware of the fact that the choice between word types and tokens can be crucial with respect to the validity of the MAL. These two papers analyse tokens, with results that do not display a decreasing trend typical for the MAL.[4]

In this paper, we focus on the relation between word length and the mean length of its constituents – i.e. syllables and characters – in Chinese. This language was thought to be an exception where the MAL does not hold at these levels in general (see Chen & Liu, 2019; Chen & Liu, 2016, 2022).

2 Word and Its Constituents in Chinese

Given the uniqueness of the Chinese script, one can consider specific methods how to evaluate the validity of the MAL if a word is taken as the construct. With the exception of erization[5] (Lin, 2007, pp. 182–189), the number of syllables in a word is equal to the number of characters the word consists of. We thus have two possible approaches, roughly corresponding to the distinction between spoken and written language. On the one hand, word length in Chinese can be measured as in other languages, i.e. in syllables, and syllables length in phonemes or pinyin letters. On the other hand, and specifically for Chinese, characters also can be used as reasonable word length units. The characters are composed of components (Wang, 2002, p. 35) which, in turn, consist of strokes (Sun, 2006, pp. 107–110).

Several studies examining Chinese words from the MAL perspective have appeared in scientific literature. The first studies which applied the law to written Chinese were published by Bohn (1998, 2002), where word length is expressed in the number of characters and character size in components. Words from a dictionary[6] were used, which implies that frequencies are not considered. The results (Bohn, 2002, p. 166) showed that the data follow the MAL.

On the contrary, studies conducted by Chen and Liu (2016, 2019, 2022) did not confirm the validity of the MAL for the same units. Therefore, they suggest measuring words directly in components (i.e. characters are omitted). This approach leads to a much higher similarity between the data and the MAL expressed by equation (1). Based on this fact, Chen and Liu (2016, 2019, 2022) believe that component, and not character, is a suitable unit for measuring word length in Chinese. Chen and Liu (2016) do not say explicitly whether they analysed types or tokens (although they provide information on texts used as research material in their paper only in terms of tokens). Two subsequent studies published by the same authors (Chen & Liu, 2019; Chen & Liu, 2022) used tokens.

To the best of our knowledge, three studies focusing on the MAL in spoken Chinese have been published so far. The above-mentioned paper by Chen and Liu (2016) examined word length measured not only in characters and their components but also in syllables. Then, syllable length was evaluated in the number of both phonemes and pinyin letters. Neither of the two possibilities corroborated the validity of the law. The study concluded that 'pinyin letter and phoneme are not the lower units of syllable in Chinese' (Chen & Liu, 2016, p. 17).

A different line of research focusing on the prosodic properties of Chinese was pursued by Ščigulinská and Schusterová (2014), where stress unit (which they define as a group of syllables separated by a pause and having at least one accent) and segment (a rhythmical unit based on speech pace) were considered constructs.[7] They were measured in the number of syllables,[8] with syllable length determined in phonemes. The mean syllable length remains more or less constant regardless of stress unit length. The relation between the rhythmic segment length and the mean syllable length abides by the MAL if an observation with the lowest frequency is omitted. It is not specified whether types or tokens were used. A follow-up study by Kovaľová and Schusterová (2016) analyzes only the relation between lengths of stress units and syllables,[9] with syllable length expressed by its duration in seconds (see also Geršić and Altmann, 1980; Rothe-Neves et al., 2017). The results revealed that the agreement with the MAL law is high in the case of spontaneous speech samples. In case of read speeches, the results vary. However, according to the authors of the study, the decreasing trend characteristic for the law is still noticeable. The recordings were segmented using Praat software,[10] which means that only tokens were analysed. On the other hand, it is questionable whether types make any sense here, as stress units are identified by pauses made by individual speakers.

Given the results (which are ambiguous at best), a question arises why the MAL does not seem to be a valid model for Chinese words measured in syllables or in characters. Two factors might be considered. Firstly, word length variability in Chinese words is limited in comparison with a majority of other languages (see, e.g., Grzybek, 2006). Chinese words are usually not longer than four syllables (or characters), with one- and two-syllable words representing the majority (Chen et al., 2015). It is questionable whether the law has ‘enough space’ to manifest itself.

Another factor which is likely to have an impact on the results is the choice between word types and tokens. For entries from dictionaries or word types from texts, longer words are – according to the MAL, as it was formulated by Menzerath (1954) and Altmann (1980) – composed of syllables which are on average shorter. The MAL is not valid for word tokens in Chinese, as can be seen in Chen and Liu (2016, 2019, 2022). This behaviour can be explained (admittedly, only speculatively for the time being) as a display of a competition between two ‘language forces’ represented by the MAL on the one hand, and by the Zipf law of brevity (Zipf, 1949; Bentz & Ferrer-i-Cancho, 2016) on the other. With tokens, frequency comes into play, and according to the brevity law, shorter tokens are used more frequently. If the law of brevity is valid also within words of particular lengths (e.g. if monosyllabic words consisting of few phonemes occur more often than monosyllabic words with more phonemes), the MAL may hold for word tokens (e.g. most of the texts analysed in Mačutek et al. 2019) but it need not (e.g. Rujević et al., 2021), depending on how strongly the law of brevity prefers shorter syllables. Also, Stave et al. (2021)[11] pointed out that unit frequency based on word tokens can have a biasing impact on results. Word types with high token frequency have a stronger influence on the mean token length than less frequent types and consequently make it difficult for the MAL to properly and fully show itself. 'Menzerath's Law is expected to be due to an intrinsic trade-off between the components and the carrier, and not to the frequency of the of usage of the specific carrier' (Stave et al. 2021: 4).

3 Language Material and Methodology

In order to shed some light on the validity of the MAL as a model for the relation between word length in the mean length of word constituents, we analyse Chinese texts considering both word types and tokens. A comparison of the obtained results can reveal whether the choice between these two approaches really has an impact. With respect to the methodology, we follow the previous studies (Bohn, 1998, 2002; Chen & Liu, 2019; Chen & Liu, 2016; Ščigulinská & Schusterová, 2014). We measure word length first in syllables, and syllable length in both phonemes and pinyin letters. Second, word length is expressed in characters, while characters are measured in components and strokes.

We conduct the experiments on two Chinese translations of the New Testament (the 27-book canon). Both versions were published online by the International Biblical Association (a non-profit organization registered in Macau, China) as a part of the Wordproject®[12] where the Bible in many other languages can be found as well. We are aware that the texts are translations which can have an impact on the results (see, e.g., Jiang & Ma, 2021, where the MAL on the level of sentence – clause – word is studied in English translations of Chinese texts). Not to mention that the translation process involves language borrowing where foreign words are transferred from a source to target language by using different methods (e.g. phonetic loans, calques). The extent to which borrowed words correspond to its original model may also influence the results. On the other hand, it offers an opportunity to compare results obtained from parallel translations of the same original text into various languages in future, which we consider an advantage.

The first of the two Chinese translations[13] is written only in pinyin transcription, it consists of 170,490 word tokens and 3,605 word types. We apply the orthographic definition of word (Wray, 2015), i.e. a word is a string of characters between spaces. Subsequently, we calculate the word length in the number of syllables which are easily identifiable and quantifiable in Chinese. Each syllable nucleus carries one of the tones. Four of them (high, rising, low, falling) are marked by different diacritics placed above vowels in pinyin. In case of a diphthong, only one of the vowels is marked, hence the occurrence of accents can distinguish between a diphthong and hiatus (if there are two neighbouring vowels in a word and both of them are marked, they form two syllables and not a diphthong). However, the neutral tone is not marked (Sun, 2006, pp. 39–40). Therefore, those sequences of two neighbouring vowels in which at least one vowel is not marked for a tone were found automatically and checked manually in pinyin texts.

A simple algorithm for an automatic determination of the number of phonemes in words was developed. The algorithm initially determines the number of pinyin letters in a word. Then, the number is adjusted for phonemes, based on the following pronunciation rules (Lin, 2007, pp. - 121–129).

(a) Post-alveolar affricates [tʂ^{h}], [tʂ], post-alveolar fricative [ʂ], and velar nasal [ŋ] are written in pinyin as digraphs <ch>, <zh>, <sh> and <ng>, respectively.
(b) Diphthongs (Lin, 2007, pp. 67–70) are written in pinyin as <ai>, <ao>, <ei> and <ou>, i.e. these sequences of vowels represent only one phoneme.
(c) If <yu> in the initial position is followed by <e> or <an>, it is replaced with [ɥ], i.e. <yue> and <yuan> are pronounced [ɥe] and [ɥɛn], respectively (Lin, 2007, p. 129).

Some differences between the numbers of pinyin letters and phonemes in words are caused by the insertion of schwa (Lin, 2007, p. 127).

(d) With the exception of <ying>, schwa is inserted when a consonant precedes <i> and velar nasal <ng>/[ŋ] directly follows it (e.g. <bing> is pronounced [bjəŋ]).
(e) Schwa is added also when <u> is preceded by a consonant other than <j>, <q>, <x> or <y>, and followed by alveolar nasal <n> (e.g. <dun> changes to [twən]).

The other translation, which is written in Chinese simplified characters[14] (166,852 word tokens and 6,111 word types), serves as language material to verify the MAL as the relation between units based on the Chinese script. Word length is measured in the number of Chinese characters the word contains. Two different units, components and strokes, are used to determine the size of a character. We thus follow the methodology of Chen and Liu (2016, 2019, 2022). Words are not separated by spaces in texts written in Chinese characters, therefore we cannot apply the orthographic definition of word. Instead, we used a Python wrapper PyNLPIR[15] developed for NLPIR ICTCLAS,[16] a well-known software for the segmentation of Chinese words. Chinese characters are easy to recognize in texts, therefore, measuring word length in this unit did not present any difficulties. Due to the lack of unified recognition of the components, segmentations of characters can vary depending on the definition used. We followed an open-source document published by Beijing Language and Culture University.[17] It provides an overview of components and number of strokes for 6,647 Chinese characters.

We are aware of the fact that the total numbers of word tokens and types in pinyin and Chinese characters are not equal. It should be noted in the first place that we used two methods of text segmentation which resulted in different totals of word tokens. In case of the New Testament in pinyin, the segmentation was carried out based on spaces between words, while the New Testament in Chinese characters was segmented by means of the Python library PyNLPIR.

Next, the total number of types is influenced by the inventory of Chinese syllables (including their diversification by tones) which is smaller compared to the inventory of Chinese characters. There are cases of words having the identical pinyin transcriptions, but different meaning distinguished by Chinese characters, e.g. *tāmen*/*他们* ('they', used for males, or the others), *她们* ('they', used for females), *它们* ('they', used for nonhuman entities). We can illustrate the drop-off in the number of word types by the difference between the translation written in Chinese characters and its version converted into pinyin by virtue of an open-source tool, a Python library pypinyin.[18] The former contains 6,111 word types while the latter 5,366 word types. The results obtained by the application of the law to the converted version are also available in the following section.

Texts written in both pinyin letters and Chinese characters were processed automatically by a Python script which was created for the purpose of this study. The data and the script with the technical documentation are freely available at https://doi.org/10.5281/zenodo.8003699.

4 Results

We used the NLREG software[19] to fit the MAL given by formula $y(x) = ax^b$ to the data. Within the context of this paper, x is word length, while $y(x)$ is the mean length of syllables or the mean size of Chinese characters. Parameter a is often replaced with the mean syllable length in monosyllabic words (see Kelih, 2010), or, more generally, with the mean size of constituents in constructs of size one. Thus, although the fit becomes slightly worse (only one free parameter remains in the formula), a solid linguistic interpretation of one of the parameters is obtained. The fit is usually considered good if $R^2 \geq 0.9$ (although sometimes even values as low as 0.75 are considered satisfactory, see, e.g., Chen & Liu, 2019; Chen & Liu, 2022). Note that this threshold value is only a rule of thumb (see Mačutek & Wimmer, 2013), and a model with a slightly worse fit does not necessarily have to be rejected.

For the sake of comparison, we present the results for both word types and tokens (see Table 1, 2, 3 and Figures 1, 2, 3, 4). Recall that we model the relation between word length and the mean length of its constituents (syllables and characters), and the means can be affected by a few relatively extreme values if the sample size is not large enough. We therefore require that the minimum frequency in each category be at least 10.[20] If this minimum is not achieved, the category with a too low frequency is pooled with its neighbour, until our criterion is met. Then, word length in the pooled category is represented by the weighted mean of lengths of all words from the category, with frequencies of particular lengths serving as the weights.[21]

The relation between word length in syllables and the mean syllable length in phonemes can be seen in Table 1 and Figure 1. The value of parameter *a* in

Table 1. Relation between word length in syllables and the mean syllable length in phonemes in the Chinese translation of the New Testament (x – word length in syllables, f_x – frequency of words of length x, $MSL(x)$ – the mean syllable length in words of length x, N – sample size, $\bar{x}$ – mean, s – standard deviation).

	Tokens		Types	
x	f_x	$MSL(x)$	f_x	$MSL(x)$
1	119,313	2.44	990	2.69
2	48,763	2.51	2328	2.63
3	1861	2.28	218	2.32
4	521	2.23	56	2.23
5	19	2.26		
5.38			13	2.25
6.08	13	2.41		
N	170,490		3605	
$\bar{x}$	1.32	2.46	1.83	2.59
s	0.51	1.48	0.64	1.76
b		−0.03		−0.11
R^2		0.2812		0.8489

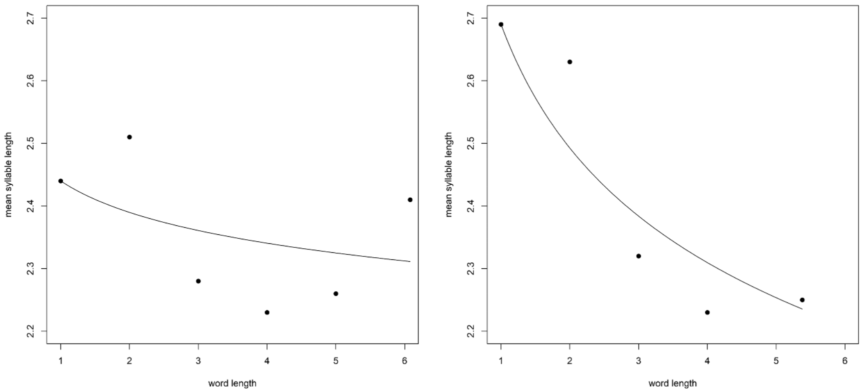

Figure 1. Word length in syllables and the mean syllable length in phonemes in the Chinese translation of the New Testament (left – tokens, right – types).

function (1) is set to be equal to the mean syllable length in monosyllabic words (i.e. $a = 2.44$ for tokens, $a = 2.69$ for types).

Obviously, the mean syllable lengths computed from word tokens do not abide by the MAL, whereas the fit for values from word types is much better. The mean syllable length decreases with the increasing word length, although the determination coefficient is slightly below the usual threshold of 0.9. If the syllable length is measured in pinyin letters (as was done by Chen & Liu, 2016), we obtain similar results ($a = 2.89$, $b = -0.07$, $R^2 = 0.2113$ for tokens, $a = 3.20$, $b = -0.17$, $R^2 = 0.8513$ for types). We note that another approach to the Chinese phonology (a different treatment of glides, see Duanmu, 2007, pp. 79–81) brings different mean lengths of syllables, but the overall trend remains unchanged (i.e. the mean syllables length decreases with the increasing word length).[22]

The relation between word length measured in the number of Chinese characters and the mean character size gives a similar picture regardless of the unit (component or stroke) used to determine the size of characters (see Table 2 and Figure 2 and 3). The mean character size in tokens behaves irregularly, but it clearly depends on word length for types. There is one exception from the decreasing tendency at length four. However, many of the longer words are words with an ambiguous segmentation (e.g. 从此以后 *cóngcǐyǐhòu* 'from now on') or fixed expressions (e.g. 自言自语*zìyánzìyǔ* ‘think aloud’, ‘talk to oneself’) which in fact consist of several shorter words. Without regard to whether the meaning of the fixed expressions can be derived from their constitutional words having their conventional meaning, the fixed expressions might behave in the same way as shorter words (e.g. if a four-syllabic word is created by

Table 2. Relation between word length and the mean character size in the Chinese translation of the New Testament (x – word length in Chinese characters, f_x – frequency of words of length x, $MCSC(x)$– the mean character size measured in components in words of length x, $MCSS(x)$– the mean character size measured in strokes in words of length x, N – sample size, $\bar{x}$ – mean, s – standard deviation).

	Tokens			Types		
x	f_x	$MCSC(x)$	$MCSS(x)$	f_x	$MCSC(x)$	$MCSS(x)$
1	111,549	2.11	7.04	1551	2.49	8.92
2	53,041	2.11	7.20	4080	2.23	7.73
3	1743	1.89	6.30	330	1.96	6.78
4	483	2.30	7.97	129	2.04	6.83
5.19	36	1.94	6.49	21	1.96	5.96
N	166,852			6111		
$\bar{x}$	1.35	2.11	7.10	1.85	2.23	7.75
s	0.52	1.47	5.26	0.64	1.72	6.08
b		−0.02	−0.01		−0.16	−0.23
R^2		0.0242	0.0046		0.8988	0.9658

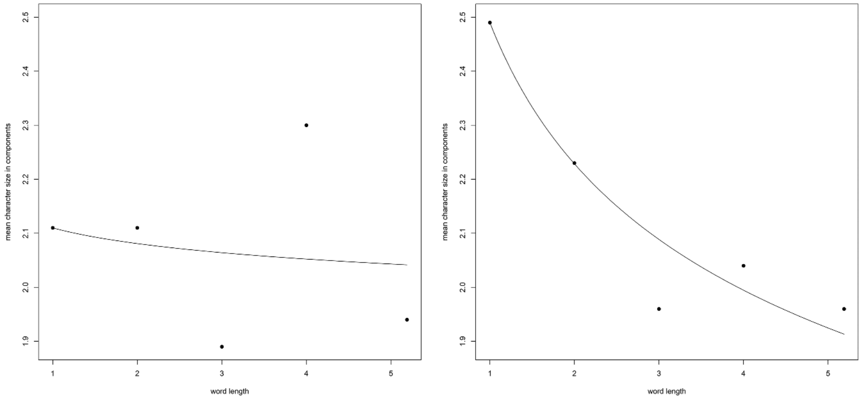

Figure 2. Word length in characters and the mean character size in components in the Chinese translation of the New Testament (left – tokens, right – types).

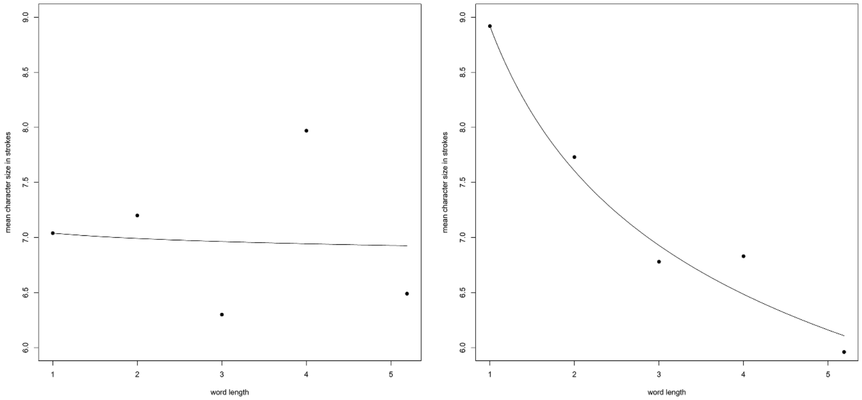

Figure 3. Word length in characters and the mean character size in strokes in the Chinese translation of the New Testament (left – tokens, right – types).

merging two two-syllabic words, it displays properties of two-syllabic words). In addition, the Bible translations contain words of a non-Chinese origin, such as proper names (e.g. 亚伯拉罕*yàbólāhǎn* ‘Abraham’) or toponyms (e.g. 耶路撒冷 *yēlùsālěng* ‘Jerusalem’). The structure of borrowed words keeps, at least partially, properties of the donor language, and there are differences in the MAL parameters among languages. Therefore, some irregularities in the mean component size of longer constructs do not have to invalidate the MAL.

The value of parameter *a* in function (1) is again set to be equal to the mean character size in words of length one.[23]

A similar trend in the results is also yielded when applying the law to a version of the New Testament written in Chinese Characters but converted into pinyin regardless of whether the syllable length is measured in phonemes (see Table 3, Figure 4) or pinyin letters ($a = 2.82$, $b = -0.08$, $R^2 = 0.6208$ for tokens, $a = 3.19$, $b = -0.15$, $R^2 = 0.8450$ for types).

Table 3. Relation between word length in syllables and the mean syllable length in phonemes in the translation of the New Testament written in Chinese characters but converted to pinyin (x – word length in syllables, f_x – frequency of words of length x, $MSL(x)$ – the mean syllable length in words of length x, N – sample size, $\bar{x}$ – mean, s – standard deviation).

	Tokens		Types	
x	f_x	$MSL(x)$	f_x	$MSL(x)$
1	111,549	2.41	851	2.69
2	53,041	2.54	4036	2.64
3	1743	2.31	329	2.36
4	483	2.31	129	2.35
5.19	36	2.23	21	2.27
N	166,852		5366	
$\bar{x}$	1.35	2.46	1.96	2.60
s	0.52	1.54	0.60	1.69
b		−0.03		−0.10
R^2		0.3520		0.8642

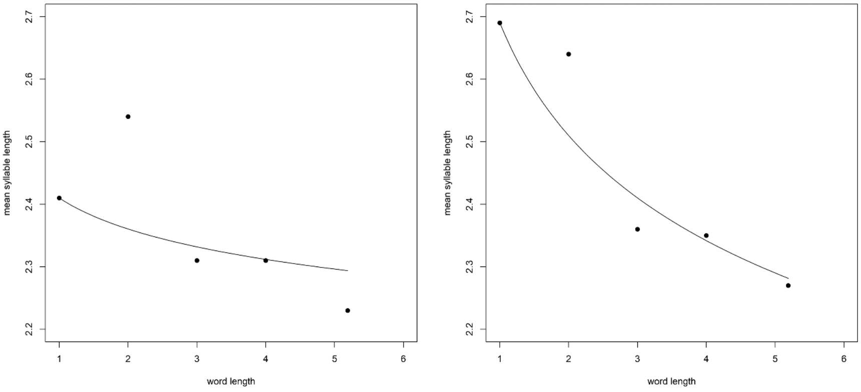

Figure 4. Word length in syllables and the mean syllable length in phonemes in the New Testament written in Chinese characters but converted to pinyin (left – tokens, right – types).

5 Conclusion

The results presented in Section 4 refute claims that word length in Chinese is exceptional with respect to the MAL. On the contrary, if word types are investigated (which is the original approach chosen by Menzerath, 1954), the mean constituent length tends to decrease with the increasing word length, regardless of units in which word length is expressed. However, data obtained from word tokens give a different picture. For these data, the relation between word length and the mean constituent length does not abide by the MAL (it is quite irregular especially if word length is measured by the number of characters).

A closer examination of the data presented in Section 4 also provides an indication that the Zipf's law of abbreviation interacts with the MAL. One can see in Tables 1 and 2 and 3that the mean constituent length in words with lengths 1, 2 and 3 is less for tokens than for types.[24] This is true for word length measured in syllables and the mean syllable length measured in phonemes as well as for word length expressed in the number of characters and character size in both components and strokes (i.e. regardless of the units in which construct and constituent size is measured). It follows that, within groups of words with particular lengths, words consisting of shorter syllables (and of simpler characters) are used more often. Thus, the Zipf law of brevity is valid also on the lower levels (shorter syllables and simpler characters are preferred) in Chinese. It influences the mean syllable length if frequencies are taken into account (i.e. if one works with word tokens, and not with types). This influence interacts with the tendency of constituents to be longer in shorter constructs, and thus it is responsible for results which do not conform to the MAL.

Several problems remain to be addressed in future. First, the MAL as a model for the relation between the length of word types and the mean syllable length is slightly worse than usual (see studies cited in Section 1). This can perhaps be explained by the text chosen (i.e. the New Testament), as it contains many non-Chinese proper names (persons, places). However, given that Chinese is a tonal language, the impact of tone should be considered. It is quite possible that in tonal languages not only syllable length but also the type of tone plays a role (e.g. tones in which pitch remains at the same level can be easier to pronounce than the ones in which the pitch level is not constant[25]). Second, tone as such requires a thorough investigation by quantitative methods before its properties can be included into the mathematical model for the MAL. Finally, the question whether the Zipf law of brevity is generally valid not only for words but also for syllables, remains open for the time being. However, Rujević et al. (2021, p. 61) report negative correlations between syllable length and syllable frequency in four Slavic languages. Also, the results from Mikros and Milička (2014), where the MAL is shown to be valid for types but not for tokens in Greek, suggest the positive answer as a reasonable conjecture.

The MAL as a mathematical model (1) for the relation between word length and the mean syllables length achieves a satisfactory fit for all languages investigated so far if word types are considered. By contrast, this relation is much more irregular for word tokens (see Section 1). Therefore, we can conclude that the MAL is a model for the lexicon (understood as a list of types or lemmas) rather than for word usage. The difference between types and tokens is self-evident for words, but it is less obvious the higher in the language unit hierarchy one moves. While it is true that, e.g., most clauses or sentences occur only once in a text or even in a corpus,[26] some very short ones (e.g. 'I don't know') can be used quite often. To the best of our knowledge, frequencies have never been considered in the context of the MAL for the 'upper neighbours' of word. The fact that clauses, sentences, etc., were always taken as they occur in language material (and not strictly types) can be one of the reasons why modelling the MAL at the syntactic level faces some still unsolved difficulties and ambiguities (see Mačutek et al., 2021, p. 66).

Notes

[1] A more general formula with an additional parameter c, $y(x) = ax^b e^{cx}$, is sometimes used, see e.g. Mačutek et al. (2019).

[2] The MAL has found its place also in research areas outside of human language, such as e.g. music (Boroda & Altmann, 1991), animal communication (Gustison et al., 2016), and genome structure (Ferrer-I-Cancho et al., 2014). The 'common denominator' of these branches of science is that they study information flow (in a very general sense).

[3] Syllable length was measured in moras, not in phonemes.

[4] In some of the papers cited in this paragraph, the mean syllable length is expressed in the number of graphemes rather than phonemes. The mean syllable length is quite similar for both choices in languages with shallow orthographies (Coulmas, 2002).

[5] Erization is an addition of the r-suffix (儿) to a syllable, e.g. 花 huā becomes 花儿 huār ('flower'). Moreover, there are a few singular exceptions of polysyllabic characters in Chinese. Qiu (2000, p. 26, 406) mentions 瓩 qiānwǎ 'kilowatt', 浬 hǎilǐ 'nautical mile', and 哩 yīnglǐ 'English mile' (none of these words occurs in our language material).

[6] Xin Han-Da cidian – Das neue Chinesisch-Deutsche Wörterbuch, 1985. Commercial Press, Beijing.

[7] In fact, one can speak about phonological words here, see e.g. Hall (1999) or Zsiga (2013, pp. 342–346). Thus, this approach can be considered a study of the MAL on the level of words, albeit from a slightly different perspective.

[8] Lengths of stress units ranged between 1 and 18 syllables while in the case of rhythmic segments between 1 and 7 syllables (Ščigulinská & Schusterová, 2014, pp. 70–72, p. 77).

[9] Kovaľová and Schusterová (2016, pp. 122–133) reported lengths of stress units between 1 and 21 syllables, similarly to Rothe-Neves et al. (2017, p. 6) who reported lengths of utterances between 2 and 29 syllables. On the other hand,

Geršić and Altmann (1980, pp. 115–123) tested the law on word lengths only up to 5 syllables.

[10] https://www.fon.hum.uva.nl/praat/ (accessed 1 June 2023).

[11] Recall that Stave et al. (2021) study the relation between word length in morphemes and the mean morpheme length in graphemes.

[12] https://www.wordproject.org/ (accessed 1 June 2023).

[13] International Biblical Association. Wordproject®: Sheng Jing: Xīnyuē Quán Shū [Holy Bible. New Testament]. Available at https://www.wordproject.org/bibles/pn/index.htm (accessed 1 June 2023).

[14] International Biblical Association. Wordproject®: 圣经. 新约全书 [Holy Bible. New Testament]. Available at https://www.wordproject.org/bibles/gb_cat/index.htm (accessed 1 June 2023).

[15] Available at https://github.com/tsroten/pynlpir (accessed 1 June 2023).

[16] Available at https://github.com/NLPIR-team/NLPIR (accessed 1 June 2023).

[17] Available at http://bcc.blcu.edu.cn/downloads/resources/%E6%B1%89%E5%AD%97%E4%BF%A1%E6%81%AF%E8%AF%8D%E5%85%B8.zip (accessed 1 June 2023).

[18] Available at https://github.com/mozillazg/python-pinyin (accessed 23 July 2023).

[19] http://www.nlreg.com (accessed June 2023)

[20] Naturally, this requirement is another rule of thumb. See e.g. Mačutek and Rovenchak (2011) and Mačutek et al. (2021) for similar, but slightly different approaches to the problem of word length categories with too low frequencies.

[21] If, e.g. we measure word length in syllables, and lengths from 1 to 5 occur more than 10 times, length 6 has frequency 12, and length 7 has frequency 1, we pool the last two lengths into one category. The weighted mean word length in this category is $\frac{12\times6+1\times7}{12+1} = 6.08$; see data in Table 1.

[22] We also obtained comparable results for the relation between word length and the mean syllables length for Pīnyīn Rìjì Duǎnwén, a diary written by Zhang Qiling (available at http://www.pinyin.info/readings/pinyin_riji_duanwen.html, accessed 1 June 2023), and for a sample containing Press reportage (text category A) and Science academic prose (text category J) from The Lancaster Corpus of Mandarin Chinese (McEnery et al., 2003). Similarly to Table 1 and Figure 1, there is a decreasing tendency of the mean syllable length, with a slight increase for the longest words.

[23] We also obtained comparable results for the relation between word length in Chinese characters and the mean character size in components and strokes, respectively, for a short story 我为什么要结婚 [Why do I want to get married] from a short story collection 黄昏里的男孩 [The boy in the dusk]) written by Yu Hua (2012), as well as for a sample containing Press reportage (text category A) and Science academic prose (text category J) from The Lancaster Corpus of Mandarin Chinese (McEnery et al., 2003).

[24] Words consisting of one, two, and three syllables make 99.7% of all word tokens in the Chinese translation of the New Testament, see Table 1.

[25] Given the wide scope of the least effort principle (see Zipf, 1949), easier-to-pronounce tones probably occur more frequently (see Zhang, 2002). Tone characteristics can also interact with other word properties, e.g. longer words can have a higher proportion of simpler tones than shorter ones.

[26] According to Berdicevskis (2021, p. 27), ‘clauses are not repeated in languages often enough to enable frequency estimates’.

Acknowledgments

The work was supported from European Regional Development Fund Project "Sinophone Borderlands – Interaction at the Edges", CZ.02.1.01/0.0/0.0/16_019/0000791 (T. Motalová), VEGA 2/0096/21 (J. Mačutek), APVV-21-0216 (J. Mačutek), and Operational Programme Integrated Infrastructure (OPII) for the project 313011BWH2: "InoCHF – Research and development in the field of innovative technologies in the management of patients with CHF", co-financed by the European Regional Development Fund (J. Mačutek).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Agentúra na Podporu Výskumu a Vývoja [APVV-21-0216]; European Regional Development Fund [CZ.02.1.01/0.0/0.0/16_019/0000791]; Operational Programme Integrated Infrastructure (OPII) [313011BWH2]; Vedecká Grantová Agentúra MŠVVaŠ SR a SAV [2/0096/21].

ORCID

Tereza Motalová http://orcid.org/0000-0002-5590-4934
Ján Mačutek http://orcid.org/0000-0003-1712-4395
Radek Čech http://orcid.org/0000-0002-4412-4588

References

Altmann, G. (1980). Prolegomena to Menzerath's law. *Glottometrika*, *1980*(2), 1–10.

Andres, J., Benešová, M., Kubáček, L., & Vrbková, J. (2012). Methodological note on the fractal analysis of texts. *Journal of Quantitative Linguistics*, *19*(1), 1–31. https://doi.org/10.1080/09296174.2011.608604

Bentz, C., & Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In C. Bentz, G. Jäger, & I. Yanovich. (Eds.), *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*. Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558

Berdicevskis, A. 2021. Successes and failures of Menzerath's law at the syntactic level. Online. In R. Čech & X. Chen (Eds.), *Proceedings of the second workshop on quantitative syntax (Quasy, SyntaxFest 2021)*, 17–32. Stroudsburg: Association for Computational Linguistics. https://aclanthology.org/2021.quasy-1.2 (Retrieved June 1, 2023).

Bohn, H. (1998). *Quantitative Untersuchungen der modernen chinesischen Sprache und Schrift*. Verlag Dr. Kovač.

Bohn, H. (2002). Untersuchungen zur chinesischen Sprache und Schrift. Online. In R. Köhler (Ed.), *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik* (pp. 127–177). Universität Trier. (Retrieved June 1, 2023 https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/deliver/index/docId/146/file/05_bohn.pdf

Boroda, M. G., & Altmann, G. (1991). Menzerath's law in musical texts. *Musikometrika*, *3*, 1–13.

Buk, S., & Rovenchak, A. (2007). Statistical parameters of Ivan Franko's novel perekhresni stežky (the Cross-Paths). In P. Grzybek & R. Köhler (Eds.), *Exact methods in the study of language and text* (pp. 39–48). De Gruyter Mouton. https://doi.org/10.1515/9783110894219.39

Chen, H., Liang, J., Liu, H. (2015). How does word length evolve in written Chinese? *PLOS ONE*, *10*(9), 1–12. https://doi.org/10.1371/journal.pone.0138567

Chen, H., & Liu, H. (2016). How to measure word length in spoken and written Chinese. *Journal of Quantitative Linguistics*, *23*(1), 5–29. https://doi.org/10.1080/09296174.2015.1071147

Chen, H., & Liu, H. 2019. A quantitative probe into the hierarchical structure of written Chinese. In X. Chen & R. Ferrer-I-Cancho (Eds.), *Proceedings of the first workshop on quantitative syntax (Quasy, SyntaxFest 2019*), 25–32. Stroudsburg: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-7904.

Chen, H., & Liu, H. (2022). Approaching language levels and registers in written Chinese with the Menzerath–Altmann Law. *Digital Scholarship in the Humanities*, *37*(4), 934–948. https://doi.org/10.1093/llc/fqab110

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

Cramer, I. M. (2005). Das Menzerathsche Gesetz. In R. Köhler, G. Altmann, & G. P. Rajmund (Eds.), *Quantitative linguistics. An international handbook* (pp. 659–688). de Gruyter.

Dinu, A., & Dinu, L. P. 2009. On the behavior of Romanian syllables related to minimum effort laws. In E. Paskaleva, S. Piperidis, M. Slavcheva, & C. Vertan (Eds.), *Proceedings of the workshop Multilingual resources, technologies and evaluation for central and Eastern European languages*, 9–13. Borovets: Association for Computational Linguistics. https://doi.org/10.5555/1859119.1859121.

Duanmu, S. (2007). *The phonology of standard Chinese*. Oxford University Press.

Ferrer-I-Cancho, R., Hernández-Fernández, A., Baixeries, J., Dębowski, Ł., & Mačutek, J. (2014). When is Menzerath-Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, *13* (6), 633–644. https://doi.org/10.1515/sagmb-2013-0034

Geršić, S., & Altmann, G. (1980). Laut – Silbe – Wort und das Menzerathsche Gesetz. In H.-W. Wodarz (Ed.), *Frankfurter phonetische Beiträge* (Vol. 3, pp. 115–123). Buske.

Grzybek, P. (1999). Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen. In B. Tošović Ed., *Die grammatischen Korrelationen. (GraLiS-1999)* (pp. 67–77). Institut für Slawistik der Karl-Franzens-Universität Graz.

Grzybek, P. (Ed.). (2006). *Contributions to the science of text and language: Word length studies and related issues*. Springer.

Gustison, M. L., Semple, S., Ferrer-I-Cancho, R., & Bergman, T. J. (2016). Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, *113*(19), E2750–E2758. https://doi.org/10.1073/pnas.1522072113

Hall, T. A. (1999). The phonological word: A review. In T. A. Hall & U. Kleinhenz (Eds), *Studies on the phonological word* (pp. 1–22). John Benjamins Publishing Company. https://doi.org/10.1075/cilt.174.02hal

Jiang, Y., & Ma, R. (2021). Does Menzerath–Altmann law hold true for translational language: Evidence from translated English literary texts. *Journal of Quantitative Linguistics*, *29*(1), 37–61. https://doi.org/10.1080/09296174.2020.1766335

Jin, H., & Liu, H. (2017). How will text size influence the length of its linguistic constituents? *Poznan Studies in Contemporary Linguistics*, *53*(2), 197–225. https://doi.org/10.1515/psicl-2017-0008

Kelih, E. (2008). Wortlänge und Vokal-/Konsonantenhäufigkeit: Evidenz aus slowenischen, makedonischen, tschechischen und russischen Paralleltexten. *Anzeiger für Slavische Philologie*, *36*, 7–27.

Kelih, E. (2010). Parameter interpretation of Menzerath law: Evidence from Serbian. In P. Grzybek, E. Kelih, & J. Mačutek (Eds.), *Text and language: Structures, functions, interrelations, quantitative perspectives* (pp. 71–78). Praesens.

Kelih, E. (2012). Systematic interrelations between grapheme frequencies and word length: Empirical evidence from Slovene. *Journal of Quantitative Linguistics*, *19*(3), 205–231. https://doi.org/10.1080/09296174.2012.685304

Kovaľová, J., & Schusterová, D. (2016). Menzerath-Altmann law – analyses of spoken Chinese. In M. Benešová (Ed.), *Text segmentation for Menzerath-Altmann law testing* (pp. 117–136). Palacký University.

Kraviarová, M., & Zimmermann, J. 2010. Menzerathov zákon v slovenskom vedeckom texte [Menzerath's law in a Slovak scientific text]. *Jazyk a kultúra [Language and Culture]1*. https://www.ff.unipo.sk/jak/1_2010/kraviarova_zimmermann.pdf (Retrieved June 1, 2023).

Kulacka, A. (2009). The necessity of the Menzerath-Altmann law. *Anglistica Wratislaviensia*, *47*, 55–60. (Retrieved June 1, 2023 https://wuwr.pl/awr/article/download/120/99/

Lin, Y.-H. (2007). *The sounds of Chinese*. Cambridge University Press.

Mačutek, J., Čech, R., & Courtin, M. 2021. The Menzerath-Altmann law in syntactic structures revisited: Combining linearity of language with dependency syntax. In R. Čech & X. Chen (Eds.), *Proceedings of the second workshop on quantitative syntax (Quasy, SyntaxFest 2021)*, 65–73. Stroudsburg: Association for Computational Linguistics. https://aclanthology.org/2021.quasy-1.6/ (Retrieved June 1, 2023).

Mačutek, J., Čech, R., & Milička, J. 2017. Menzerath-Altmann law in syntactic dependency structure. Online. In S. Montemagni & J. Nivre (Eds.), *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)*, 100–107. Linköping: Linköping University Electronic Press. https://aclanthology.org/W17-6513.pdf (Retrieved) June 1, 2023).

Mačutek, J., Chromý, J., & Koščová, M. (2019). Menzerath-Altmann law and prothetic/v/in spoken Czech. *Journal of Quantitative Linguistics*, *26*(1), 66–80. https://doi.org/10.1080/09296174.2018.1424493

Mačutek, J., & Mikros, G. K. (2015). Menzerath-Altmann law for word length motifs. In G. K. Mikros & J. Mačutek (Eds.), *Sequences in language and text* (pp. 125–132). de Gruyter. https://doi.org/10.1515/9783110362879-009

Mačutek, J., & Rovenchak, A. A. (2011). Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In E. Kelih, V. Levickij, & Y. Matskulyak (Eds.), *Issues in quantitative linguistics* (Vol. 2, pp. 136–147). RAM-Verlag.

Mačutek, J., & Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, *20*(3), 227–240. https://doi.org/10.1080/09296174.2013.799912

McEnery, A., Xiao, Z., & Mo, L. (2003). Aspect marking in English and Chinese: Using the Lancaster corpus of mandarin Chinese for contrastive language study. *Literary and Linguistic Computing*, *18*(4), 361–378. https://doi.org/10.1093/llc/18.4.361

Menzerath, P. (1954). *Die Architektonik des deutschen Wortschatzes*. Dümmler.

Menzerath, P., & de Oleza, J. M. (1928). *Spanische Lautdauer: Eine experimentelle Untersuchung*. de Gruyter. https://doi.org/10.1515/9783111729008

Mikros, G., & Milička, J. (2014). Distribution of the Menzerath's law on the syllable level in Greek texts. In G. Altmann, R. Čech, J. Mačutek, & L. Uhlířová (Eds.), *Empirical approaches to text and language analysis* (pp. 181–189). RAM-Verlag.

Milička, J. (2014). Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, *21*(2), 85–99. https://doi.org/10.1080/09296174.2014.882187

Pelegrinová, K., Mačutek, J., & Čech, R. (2021). The Menzerath-Altmann law as the relation between lengths of words and morphemes in Czech. *Jazykovedný časopis [Journal of Linguistics]*, *72*(2), 405–414. https://doi.org/10.2478/jazcas-2021-0037

Qiu, X. (2000). *Chinese writing*. The Society for the Study of Early China.

Rothe-Neves, R., Marques Bernardo, B., & Espesser, R. (2017). Shortening tendency for syllable duration in Brazilian Portuguese utterances. *Journal of Quantitative Linguistics*, *25*(2), 156–167. https://doi.org/10.1080/09296174.2017.1360172

Rovenchak, A. (2015). Quantitative studies in the porpus of nko periodicals. In A. Tuzzi, M. Benešová, & J. Mačutek (Eds), *Recent contributions to quantitative linguistics* (pp. 125–138). de Gruyter. https://doi.org/10.1515/9783110420296-012

Rujević, B., Kaplar, M., Kaplar, S., Stanković, R., Obradović, I., & Mačutek, J. (2021). Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian. In A. Pawłowski, J. Mačutek, S. Embleton, & G. Mikros (Eds.), Language and text: Data, models, information and applications *(Current issues in linguistic theory* (Vol. 356, pp. 55–67). John Benjamins. https://doi.org/10.1075/cilt.356.04ruj

Ščigulinská, J., & Schusterová, D. (2014). *An application of the Menzerath–Altmann law to contemporary spoken Chinese*. Palacký University Olomouc.

Stave, M., Paschen, L., Pellegrino, F., & Seifart, F. (2021). Optimization of morpheme length: A cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard*, *7*(s3). https://doi.org/10.1515/lingvan-2019-0076

Sun, C. (2006). *Chinese: A linguistic introduction*. Cambridge University Press.

Wang, N. (2002). *汉字构型学讲座 [The lecture on the composition and formation of Chinese characters]*. Shanghai Educational Publishing House.

Wang, L., & Čech, R. (2016). The impact of code-switching on the Menzerath-Altmann law. *Glottometrics*, *35*, 22–27.

Wray, A. (2015). Why are we so sure we know what a word is? In J. R. Taylor (Ed), *The oxford handbook of the word* (pp. 725–750). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641604.013.032

Yu, H. (2012). 黄昏里的男孩 [The boy in the dusk]. Beijing: 作家出版社 [Writers Publishing House].

Zhang, J. (2002). *Effects of duration and sonority on contour tone distribution. A typological survey and formal analysis*. Routledge.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press.

Zsiga, E. C. (2013). *The sounds of language: An introduction to phonetics and phonology*. Wiley.