Palacký University Olomouc
Faculty of Arts
Department of General Linguistics

Menzerath-Altmann Law in Chinese

Tereza Motalová

Supervisor

doc. Mgr. Radek Čech, PhD.

Department of Czech Language
Faculty of Arts
University of Ostrava

PhD. Thesis Summary | 2022

Contents

Introduction

The Menzerath-Altmann law predicts that lengths of two language units of different hierarchical levels – a hierarchical higher construct and a hierarchical lower constituent – are negatively correlated. While the length of the construct lengthens, the length of the constituent shortens on average. Deviations from this general tendency occur but do not undermine the law's validity. The law is stochastic and deviations are even expected "as a consequence of the stochastic nature of the language mechanism" (Köhler, 2012, p. 175). Nowadays, the Menzerath-Altmann law is perceived as a general mechanism that maintains equilibrium in cognitive workload by regulating information flow.

Over the last four decades, the law has been corroborated when applied to various language units and language material. However, particular language units (e.g. word) are drawing more attention from researchers than others (e.g. phrase). Moreover, only one pair of the construct and its constituent is usually tested (e.g. sentence and clause accordingly) despite a unit possibly occupying different hierarchical positions (e.g. clause becoming the construct). It is also generally presumed that the negative correlation between unit lengths appears when immediate hierarchical neighbouring units are analysed. This poses a question of unit choice and unit neighbourhood which are not always apparent (e.g. clause and word vs clause and phrase).

We aim to address these challenges within the thesis. Firstly, we test the law throughout a hierarchy of chosen language units in Chinese, including the phrase that has generally been drawing less attention. The tested hierarchy consists of a sentence, clause, phrase, word, character/syllable, component/sound and stroke. It allows us to analyse how the units behave in relation to the law when their hierarchical position changes from the constituent to the construct (except for the sentence, syllable, component and stroke). Secondly, we apply the law to various unit combinations to shed light on the unit neighbourhood. Thirdly, considering the law as a general mechanism maintaining equilibrium in cognitive workload, we evaluate construct and constituent lengths with regard to limits of short-term memory represented by Miller's 'magical number plus or minus two' (1956). Fourthly, relationships between lengths of the language units mentioned above are tested on Chinese language material. Even though studies focusing on Chinese already applied the law to a hierarchy of language units, they left the phrase level out of the analysis (Bohn, 1998, 2002; Chen and Liu, 2019, 2022). Hence, including the phrase into our unit hierarchy while using its different determinations will provide valuable insights into its behaviour towards the law and other units in Chinese. Finally, Chen and Liu (2016, 2019, 2022) yielded that the law does not come into force when applied to the word being the construct and the Chinese character being its constituent. The results indicate that the law competes against the word length distribution in Chinese. The prevalence of one- and two-character/syllable words (e.g. Chen, Liang and Liu, 2015) might not provide the law with enough 'space' to manifest itself. The thesis aims to examine whether other factors influence the results or the specific word length distribution in Chinese can be regarded as the boundary conditions for the law.

1 Menzerath-Altmann law

In 1954, Menzerath published a work where he corroborated a particular lawful relationship – "[d]ie relative Lautzahl nimmt mit steigender Silbenzahl ab"[1] (Menzerath, 1954, p. 100) – for more than 20k German words. He generalized the observartion as follows "je größer das Ganze, um so kleiner die Teile!"[2] (Menzerath, 1954, p. 101) and interpreted it as a result of economy rules. Altmann (1980) reformulated Menzerath's findings while using general terms common in linguistics – a construct (being a hierarchically higher unit and corresponding to Menzerath's whole) and a component or constituent (being a lower unit in the hierarchy and corresponding to the part in Menzerath's view). His first reformulation was as follows: "[t]he longer a language construct the shorter its components (constituents)" (Altmann, 1980, p. 1). Based on the verbal expression, Altmann suggested the following equation:

$$y = ae^{-cx}, \tag{1}$$

where the independent variable x represents a construct length, the dependent variable y is a constituent length related to the given construct, and a, c are parameters.

Since the first equation (1) only expresses a monotonic constant decrease of the constituent length which might not always hold true, Altmann, therefore, changed the first verbal expression to "[t]he length of the components is a function of the length of language constructs" (Altmann, 1980, p. 3) and adjusted the equation by addition of a parameter b responsible for "an inverse proportionality of the decrease rate to the construct length" (Altmann, 1980, p. 3):

$$y = ax^{b}e^{-cx}. \tag{2}$$

The last formula is obtained when $c = 0$ (Altmann, 1980, p. 3), i.e.

$$y = ax^{b}. \tag{3}$$

The parameter a is usually described as a value on the y-axis where a fitting curve starts if the model (3) is applied. The value approximately equals the mean size of constituents belonging to a one-constituent construct. Köhler (1982, p. 110) demonstrated the equality by inserting the construct length $x_1 = 1$ into the formula (3), i.e. $y = ax^{b}$, resulting in $y_1 = a{x_1}^{b} = a1^{b} = a$. Therefore, the parameter a can be replaced with the empirical value of y_1 in this model, i.e. $y = y_1x^{b}$ (e.g. Köhler, 1984, p. 180; Cramer, 2005b, p. 50; Kelih, 2010, p. 75).

The parameter b shows a shortening tendency, i.e. a degree to which the length of the constituent (hypothetically) shortens while the length of the construct lengthens (e.g. Köhler,

[1] "The relative number of sounds decreases as the number of syllables increases" (Menzerath, 1954, p. 100), translated by the author.

[2] "the greater the whole, the smaller the parts!" (Menzerath, 1954, p. 101), translated by the author.

1984, p. 180; Kelih, 2010, p. 71). The greater its negative value is with respect to the model (3), the steeper the decrease of a curve depicting the function y is (e.g. Hřebíček, 2002b, pp. 55-56).

The relation between both the parameters has been under discussion since the mathematical formalisation of the law. Teupenhayn and Altmann addressed that "the steepness of the curve is a function of a, i.e. the absolute value of b is proportionate to $[a]$" (1984, p. 129). Altmann and Schwibbe (1989, p. 43 and pp. 57-58) expected that the higher the starting value of a fitting curve, the steeper the slope of the curve, hence, values of both the parameters should be correlated. The negative correlation, i.e. with increasing value of the a parameter, the value of the b parameter decreases, was confirmed by Hammerl and Sambor (1993), Hou et al. (2019a, p. 36) or Jiang and Jiang (2022, pp. 10-11).

The parameter c is the least known parameter with respect to linguistic interpretation, and it has been addressed to a minimal extent in comparison to a and b (to our best knowledge).

The model (2), i.e. $y = ax^b e^{-cx}$, where $b \neq 0,\ c \neq 0$, is considered a general form of the law (e.g. Roukk, 2007, p. 605). On the one hand, it contains the parameter c without its solid linguistic interpretation. On the other hand, it enables to reflect a tendency which contradicts the original menzerathian assumption of the decrease in constituent lengths, i.e. a tendency of constituent lengths to increase simultaneously with the lengths of the construct (e.g. Mačutek, Chromý and Koščová, 2018, p. 2). The model (3), i.e. $y = ax^b$, is regarded as an alternative to the general model (2), i.e. $y = ax^b e^{-cx}$, where $c = 0$. It includes only two parameters, which makes it easier to interpret and preferred over the general one. The model "has turned out to be the most commonly used 'standard form' for linguistic purposes" (Grzybek and Stadlober, 2007, p. 205), and it has become sufficient in comparison with the model (2) (Köhler, 1982, p. 106).

The law has been corroborated by a number of studies which applied the law to various language materials and language units. Corroboration of the law also comes from fields across the borders of linguistics, such as musicology (Boroda and Altmann, 1991) or biology, where the law was tested on proteins (Shahzad, Mittenthal and Caetano-Anollés, 2015), genes and genomes (e.g. Sun and Caetano-Anollés, 2021), or animal communication (e.g. James et al., 2021, Valente et al., 2021).[3] However, there are also results which rejected the law (e.g. in Roukk, 2007; Buk and Rovenchak, 2008; Buk, 2014; Hou et al., 2017).

As Köhler (2012, p. 175) pointed out, the stochastic laws – which the Menzerath-Altmann law is believed to be – "include in their predictions the deviations which are to be expected as a consequence of the stochastic nature of the language mechanism concerned" (Köhler, 2012, p. 175). The deviations from the Menzerath-Altmann law were already anticipated by Altmann (1980, p. 5) and they are not considered to be a reason for its rejection – as a flight of an aeroplane being beyond boundary conditions for validity of the gravity law (Teupenhayn and Altmann, 1984, p. 130). For example, the law might manifest itself only when the construct length exceeds a specific limit – if the construct is short enough, its constituents cannot or do not need to be shortened (Schwibbe, 1984, p. 162; Kułacka, 2008, p. 174). A limit imposed by a text size was suggested by Čech and Mačutek (2021, p. 8) based on results

[3] Overviews available in Semple, Ferrer-i-Cancho and Gustison (2021, p. 6) and Torre, Dębowski and Hernández-Fernández (2021, p. 2).

obtained from a poem whose length of 94 word types was probably too short for the mechanism of law to be launched. Moreover, language is viewed as a self-organising dynamic system involving cooperative and competitive processes (Köhler, 2012, p. 170). The existence of 'forces' overlapping or counteracting the Menzerath-Altmann law has been mentioned. Such examples can be text production under abnormal conditions or an author pursuing a specific goal and consequently obeying other laws which override the Menzerath-Altmann law (Teupenhayn and Altmann, 1984, pp. 129-130; Čech and Mačutek, 2021, p. 12).

However, the validity of the law does not face only the interaction of different – known and unknown – processes or laws but also practical and theoretical challenges which relate to sampling, interrelation of linguistic properties, units of measurement or evaluation of results (cf. Grotjahn and Altmann, 1993). As regards the sampling, one of the discussed issues is the degree of heterogeneity of a language material (Almann, 1992, p. 287) which can lead to disagreement between the model and data. Since a text is produced in a particular context, a combination of texts can result in a mixed – heterogeneous – sample which some researchers prefer to avoid (e.g. Altmann, 1992, p. 291; Wimmer et al., 2003, p. 89). On the other hand, a mixed sample can also cause the mechanism to be amplified more than in individual texts (Čech, 2020, pp. 26-28). The interrelation of linguistic properties relates to the frequency of usage, i.e. unit tokens. However, there is another approach to consider (e.g. in Altmann, 1992, p. 291) when only different forms of the unit, i.e. its types, are analysed (e.g. different word forms from a text or lemmas from dictionaries).[4] This approach instead reflects a language structural property. The frequency of usage (i.e. unit tokens) closely relates to Zipf's law of abbreviation (or Brevity law) which describes the negative correlation between the unit lengths and their frequencies. Suppose the Brevity law is taken into account. In that case, the frequencies can be biased towards shorter units in a sample which applies not only to the construct but also to the constituent and, consequently, imposes double limits on the Menzerarth-Altmann law to manifest itself fully (e.g. Pelegrinová, Mačutek and Čech, 2021; Stave et al., 2021). The Menzerath-Altmann law operates with the concept of the construct and constituent standing for units of measurement. As Altmann (1983; Altmann and Schwibbe, 1989, pp. 46-48; Cramer, 2005a, pp. 633-634) pointed out, the negative correlation between lengths of the construct and the constituent only emerges if the immediately adjacent units are tested, or in other words, the levels are not skipped. However, determining individual linguistic levels and their language units is not always apparent and unambiguous. The last issue to be discussed here is the evaluation of results. The goodness-of-it between the model and data is commonly evaluated by the coefficient of determination R^2 which reflects the degree of agreement between empirical and theoretical values (Kelih, 2008, p. 17). Its value ranges from 0 to 1. The higher the value, the better fit between a model and data. However, researchers do not agree on a minimum threshold for the law's corroboration when interpreting obtained results. The lack of consensus on the threshold blurs an overall picture regarding the scope of the law's validity.

[4] We use the term 'types' to denote both – not only different word forms from a text but also basic forms of words which correspond to entries in dictionaries, i.e. lemmas (Taylor, 2015, pp. 2-3).

2 Menzerath-Altmann law in Chinese

The chapter summarises findings yielded by studies on Chinese according to the constructs tested by the thesis, i.e. sentence, clause, syntactic phrase, word and character (for their overview, see Table 1). When summarising the studies, we follow interpretations provided by authors. If the coefficient of determination R^2 is used for the evaluation of the goodness-of-fit between models and data, we additionally review the results in the light of the standard followed by the thesis, i.e. the law is corroborated when $R^2 \geq 0.90$ (Mačutek and Wimmer, 2013, p. 233).

Table 1. Overview of linguistic levels analysed by the thesis and studies on Chinese.

Construct	Direct constituent	Sub-constituent	Studies on Chinese
Sentence	Clause	Word	Bohn (1998, 2002); Wang and Čech (2016); Hou et al. (2017); Jin and Liu (2017); Chen (2018); Chen and Liu (2019, 2022); Berdicevskis (2021)*; Sun and Shao (2021)
	Sentential phrase	Word	–
	Clause	Clausal phrase	Berdicevskis (2021)*
	Clause	LDS	–
Clause	Word	Character/syllable	Bohn (1998, 2002); Hou et al., (2019a, 2019b); Berdicevskis (2021)*; Chen and Liu (2022)
	Clausal phrase	Word	Berdicevskis (2021)*
	LDS	Word	–
Sentential phrase	Word	Character/syllable	–
Clausal phrase	Word	Character/syllable	Berdicevskis (2021)
LDS	Word	Character/syllable	–
Word type	Character	Component	Bohn (1998, 2002)
	Character	Stroke	–
	Syllable	Sound	–
Word token	Character	Component	Motalová and Matoušková (2014); Chen and Liu (2016, 2019, 2022);
	Character	Stroke	Chen and Liu (2019, 2022);
	Syllable	Sound	Chen and Liu (2016)
Character type	Component	Stroke	Bohn (1998, 2002)
Character token	Component	Stroke	Motalová et al. (2013); Motalová and Matoušková (2014); Matoušková and Motalová (2015); Matoušková (2016)

2.1 The sentence in Chinese

Studies focusing on Chinese combined the sentence with the clause and word. The sentence in Chinese was usually determined as a segment between punctuation marks, i.e. a full stop, a question mark, an exclamation mark (Bohn, 1998, 2002; Hou et al., 2017) or an ellipsis[5] (Jin and Liu, 2017). Some authors relied on the sentence determination provided by an annotation scheme of language material (Wang and Čech, 2016; Hou et al., 2017; Chen, 2018; Chen and Liu, 2019, 2022; Berdicevskis, 2021) or available software (Sun and Shao, 2021).

Since tested samples usually lacked annotation of the clause, there is an apparent consensus among studies to prefer particular punctuation marks as indicators of clausal borders. Authors usually chose a comma (Chen and Liu, 2022) together with a semicolon (Hou et al., 2017; Chen, 2018; Chen and Liu, 2019) and a colon (Bohn, 1998, 2002; Jin and Liu, 2017). Sun and Shao (2021) used all these marks and extended the selection by the ellipsis. Jin and Liu (2017), Chen (2018), and Chen and Liu (2019, 2022) explained this preferred determination by a rough correspondence between the Chinese clause and a segment inserted into two punctuation marks while referring to Luke (2006). Wang and Čech (2016) and Berdicevskis (2021) are the only studies which did not use punctuation to identify the clause in Chinese. While the former study determined the clause as a sequence of words connected through syntactic relations, which includes a subject and a predicate, Berdicevskis (2021) relied on the annotation of language material.

Lastly, the word was mainly determined by software (Hou et al., 2017; Jin and Liu, 2017; Sun and Shao, 2021)[6] or authors relied on the annotation or word segmentation of language material under analysis (Bohn, 1998, 2002; Hou et al., 2017; Chen, 2018; Chen and Liu, 2019, 2022; Berdicevskis, 2021). Wang and Čech (2016) did not specify any detail concerning the word determination, but the description of language material and methodology implies that they also used the annotation.

We start with studies which corroborated the hypothesis mentioned above with respect to interpretations provided by authors. Bohn (1998, 2002) did not reject the hypothesis when testing a corpus of news (the coefficient of determination R^2 reached the standard of $R^2 \geq 0.90$). Wang and Čech (2016) concluded that samples of Chinese monolingual sentences and Chinese-English code-switching sentences follow the menzerathian tendency despite some deviations in the latter sample (nevertheless, R^2 only of the former sample is in accord with the standard of $R^2 \geq 0.90$). Hou et al. (2017) corroborated the hypothesis for a) a corpus of news broadcasting and b) text collections of written text types from the Lancaster Corpus of Mandarin Chinese (LCMC, McEnery, Xiao and Mo, 2003). When evaluating their results, $R^2 \geq 0.90$ is reached only in the case of news broadcasting and four[7] out of 11 LCMC text collections. Jin and Liu (2017) showed corroborating results of four corpora of different text types – microblogs,

[5] The ellipsis strictly denotes the punctuation mark composed of three or six dots.

[6] Hou et al. (2017) and Jin and Liu (2017) used the Chinese Lexical Analysis System ICTCLAS (Institute of Computing Technology of Chinese Academy of Science, n.d.), while Sun and Shao (2021) used the Language Technology Platform developed by Harbin Institute of Technology (Che, Li and Liu, 2010).

[7] a) news reportage, b) news editorials, c) skills, trades and hobbies, and d) academic prose.

news, prose and fiction (nonetheless, only the microblogs meet $R^2 \geq 0.90$). Chen (2018) and Chen and Liu (2019, 2022[8]) also tested LCMC, and, in the author's view, the sample did not reject the hypothesis. However, none of these studies showed R^2 reaching $R^2 \geq 0.90$). Berdicevskis (2021) confirmed a negative correlation between the units on this level for a mixed sample of UD treebanks (based on Spearman's rank correlation coefficient). Finally, Sun and Shao (2021) did not reject the hypothesis for five corpora of news, novels, prose, scripts and textbooks (R^2 reaches the standard of $R^2 \geq 0.90$ in novels, prose and scripts while in textbooks is slightly below, i.e. $R^2 = 0.8848$).

Cases which did not pass the criteria for the law's corroboration in the view of authors were reported only in two studies.[9] Firstly, when Bohn (1998, 2002) tested an individual text and secondly when Hou et al. (2017) tested corpora of texts representing informal, spontaneous language (sitcom conversations and TV talk shows) and fictional and humorous texts from LCMC.

2.2 The clause in Chinese

The clause in the position of the construct was mostly combined with the word (being its direct constituent) and the Chinese character (being its indirect constituent). The punctuation marks being borders for the clause prevailed. Authors determined the clause by using a comma (Chen and Liu, 2022) in combination with a semicolon (Chen and Liu, 2019) and a colon (Bohn, 1998, 2002; Hou et al., 2019a, 2019b). Only Berdicevskis (2021) deployed an annotation of language material. The word was identified by means of a program for word segmentation (Hou et al., 2019a, 2019b)[10], or language materials were already annotated or segmented into words (Bohn, 1998, 2002; Chen and Liu, 2019, 2022; Hou et al. 2019a; Berdicevskis, 2021). Lastly, the word length was measured in the number of Chinese characters, which roughly corresponds to the number of syllables except for erisation (Bohn, 1998, 2002; Hou et al. 2019a, 2019b; Berdicevskis, 2021; Chen and Liu, 2022).

Let us summarise the achieved results according to the interpretations of the authors. The hypothesis mentioned above was corroborated by Bohn (1998, 2002), who tested an individual text and a sample of news. However, the coefficient of determination R^2 of the text was below the standard of $R^2 \geq 0.90$, i.e. $R^2 = 0.8789$, and the sample did not even reach or approximate it. Hou et al. (2019a) did not reject the hypothesis for samples of news broadcasting, sitcom conversations and TV talk shows. When reviewing their results, none of the values of R^2 follow the standard of $R^2 \geq 0.90$. However, Hou et al. (2019a) fitted the data with a linear model of the law.[11] When Hou et al. (2019b) refitted the data with the complete model, only the

[8] The LCMC sample tested by Chen and Liu (2022) contained two text collections of press reportages and academic prose.

[9] Following the interpretation of the authors, empirically gained data showed an increasing tendency of mean clause lengths contradicting the law, or a value of the coefficient of determination R^2 was lower than 0.70.

[10] Hou et al. (2019a, 2019b) used the Chinese Lexical Analysis System ICTCLAS (Institute of Computing Technology of Chinese Academy of Science, n.d.).

[11] $y = bx + \ln(a)$

sitcom conversations would not corroborate the law concerning $R^2 \geq 0.90$. The law also applied to the Lancaster Corpus of Mandarin Chinese (LCMC) by Hou et al. (2019a), who tested its five text collections, and by Chen and Liu (2022), who tested a sample containing its two text collections (nevertheless, R^2 did not reach the standard of $R^2 \geq 0.90$ in any of these studies). Berdicevskis (2021) applied the law to mixed UD treebanks and confirmed neither a negative nor a positive correlation.[12] The only language material which was reported not to be in line with the law was a mixture of news broadcasting, sitcom conversations and TV talk shows (Hou et al., 2019a, based on R^2).[13]

Lastly, Chen and Liu (2019, 2022) tested an alternative to the word constituent. The authors left Chinese characters out and measured the word in subparts of the Chinese characters, i.e. components. Both the studies applied the law to LCMC (not further specified in 2019, while to a sample of two text collections in 2022) and achieved similar results as in the case of the Chinese characters. Nevertheless, none of the values of R^2 corroborate the law when taking $R^2 \geq 0.90$ into account.[14]

2.3 The syntactic phrase in Chinese

The syntactic phrase in the construct position was analysed only by Berdicevskis (2021) within his study of 78 languages, including Chinese. The author chose the word and the grapheme as its direct and indirect constituents respectively. Berdicevskis (2021) operationalised the phrase as a whole subtree which directly depends on a predicate (proposed by Mačutek, Čech and Milička, 2017) and is measured in the number of words belonging to it. An annotation scheme of language material provided the word determination, and the word length was expressed as a sum of its graphemes. The results showed that none of the correlations was confirmed, based on Spearman's rank correlation coefficient.[15] Otherwise, no other studies applied the law to the phrase in Chinese. Chen and Liu (2019, 2022) explained its exclusion by a problematic determination. In addition, the authors concluded with regard to their results that the word can be the direct constituent of the clause. However, when reviewing the results by optics of the standard followed by this work ($R^2 \geq 0.90$), the law would not be

[12] Based on data available at Github (AleksandrsBerdicevskis/menzerath/results_means_clause_50.tsv, 2021). If an absolute value of a correlation coefficient ranged in the interval of $\langle 0.30; 0.70 \rangle$ and the *p*-value was greater than 0.05, none of the correlations was confirmed, as in the case of Chinese.

[13] The authors considered their results tolerable if $0.70 < R^2 < 0.90$ (Hou et al., 2019a, p. 29). However, R^2 of this sample was extremely below the lower threshold.

[14] Chen and Liu (2019, 2022) used the same model to fit the data ($y = ax^b e^{-cx}$) and the coefficient of determination R^2 obtained from the triplet of the clause, word and component reached the value of 0.7657 (2019, 2022) and from the triplet of the clause, word and character the value of 0.7477 (this combination was tested only in Chen and Liu, 2022).

[15] The data for this linguistic level is available at (AleksandrsBerdicevskis/menzerath/results_means_phrasewordgrapheme_50.tsv (2021). The absolute value of a correlation coefficient is in the interval of $\langle 0.30; 0.70 \rangle$ and the *p*-value is greater than 0.05.

corroborated. Similarly, Sun and Shao (2021) added that the phrase might correspond to the clause.

2.4 The word in Chinese

Only a few studies applied the law to the word in Chinese. The choice of the word direct constituent is usually straightforward – the number of Chinese characters in a word roughly equals the number of syllables. However, the choice of the indirect constituent depends on researchers giving a preference either to phonetic transcriptions using alphabetic characters (i.e. phonemes or letters) or to the Chinese writing system (i.e. components or strokes).

The word was directly determined based on a dictionary under analysis (Bohn, 1998, 2002) and annotation of a corpus (Chen and Liu, 2019, 2022). Motalová and Matoušková (2014) carried out the word segmentation manually while applying syntactic rules by Švarný and Uher (2001), and Chen and Liu (2016) segmented their sample into words by software[16].

To our best knowledge, the combination of the word, syllable and phoneme (or grapheme) was tested only by Chen and Liu (2016). As mentioned above, the number of syllables equals the number of Chinese characters. Hence, the authors just used the Chinese characters for the syllable count. In the case of the phoneme, a pronunciation list for Chinese characters was used (without a reference). The grapheme was determined as a Latin letter of pinyin transcription.[17] The study analysed word tokens from a corpus of dialogic text and did not corroborate the hypothesis either for the phoneme or the grapheme.[18]

The word was measured in Chinese characters when giving preference to the Chinese writing system. Regarding the sub-constituents, the number of strokes in each Chinese character is immutable, whereas the number of the components depends on a chosen approach. Bohn (1998, 2002) decomposed the Chinese characters based on a modified list of components published by Stalph (1989) and Chen and Liu (2016, 2019, 2022) based on the CJK Unified Ideographs of Unicode (Laboratory for Chinese Character Research and Application, n.d.) which includes sums of the components and the strokes for more than 20k Chinese characters. Motalová and Matoušková (2014) introduced their approach to the components (for more detail, see Chapter 2.5).

The law was corroborated for the triplet of word, character and component only when Bohn (1998, 2002) tested word types from a dictionary (the coefficient of the determination R^2 agreed with $R^2 \geq 0.90$). The analyses of word tokens achieved opposite results when Motalová and Matoušková (2014) analysed an individual text, Chen and Liu (2016) a prose text corpus and Chen and Liu (2019, 2022) The Lancaster Corpus of Mandarin Chinese. Chen and Liu (2019, 2022) also applied the law to the triplet of the word, character, and stroke, but the word tokens yielded similar unsatisfactory results. Since Chen and Liu corroborated the hypothesis neither for the

[16] I.e. the Chinese Lexical Analysis System ICTCLAS (Institute of Computing Technology of Chinese Academy of Science, n.d.).

[17] The authors converted the Chinese characters into pinyin by a Java library Pinyin4j (Pinyin4j, n.d.).

[18] Specified by authors in their later study (Chen and Liu, 2022, p. 4).

component (2016, 2019, 2022) nor the stroke (2019, 2022), the authors decided to leave the Chinese character out of the unit hierarchy and to measure the word directly in components and indirectly in strokes. In their view, the results corroborated the law. R^2 obtained from The Lancaster Corpus of Mandarin Chinese was only slightly below the standard, i.e. $R^2 = 0.8982$ (Chen and Liu, 2019, 2022). However, R^2 in Chen and Liu (2016) was provided only illustratively for three out of 20 texts and only one of them would reach the standard of $R^2 \geq 0.90$.

2.5 The character in Chinese

The last language unit being the construct tested within this thesis is a basic unit of Chinese writing systems – the character – being measured directly in its components and indirectly in its strokes. In general, the character always occupies a graphic field of the same size without regard to its complexity. The Chinese characters have been analysed so far by Bohn (1998, 2002), Motalová et al. (2013), Motalová and Matoušková (2014), Matoušková and Motalová (2015) and Matoušková (2016).

The component is generally considered a structural unit smaller than the character but greater than the stroke. As for its precise determination, Bohn (1998, 2002) opted for a list of components of kanji characters compiled by Stalph (1989) with slight modifications applied. The rest of the studies adopted an alternative graphical approach which determined the component as a stroke or a group of strokes connected to each other while being separated from other groups or strokes (Motalová et al., 2013; Motalová and Matoušková, 2014; Matoušková and Motalová, 2015; Matoušková; 2016). Regarding the strokes, each character in both languages has its immutable inventory.

In the case of types, the hypothesis was corroborated for simplified Chinese characters from a computer standard GB 2312-80 (Bohn, 1998, 2002). The coefficient of the determination R^2 followed the standard of $R^2 \geq 0.90$ in both the studies. The same results were achieved for the tokens while testing the simplified Chinese characters (Motalová et al., 2013; Motalová and Matoušková, 2014; Matoušková and Motalová, 2015, with one exception when goodness-of-fit did not reach $R^2 \geq 0.90$; Matoušková, 2016) as well as the traditional Chinese characters (Motalová and Matoušková, 2014; Matoušková, 2016; satisfying $R^2 \geq 0.90$). The corroboration of the hypothesis did not come only from one translation of the poem 'The Raven' (Matoušková and Motalová, 2015).

3 Methodology

3.1 Language material

The choice of the language material was motivated by the possibility of analysing all chosen language units, including those which are determined based on dependency syntax. Therefore, we primarily opted for a material released by the Universal Dependencies (UD) project (e.g. Nivre et al., 2020; de Marneffe et al., 2021) which builds on dependency grammar and provides treebanks for various languages while utilising a unified morphosyntactic annotation (Zeman et al., 2021). We use three UD treebanks for Chinese – Chinese-HK UD treebank (Wong et al., 2017), Chinese Parallel Universal Dependency (Zeman et al., 2017) and UD Chinese GSDSimp (UD Chinese GSDSimp, 2021).[19] When the law is applied to the word and character level, we additionally opted for The Lancaster Corpus of Mandarin Chinese (McEnery, Xiao and Mo, 2003). For an overview of the samples, see Table 2.

Table 2. Overview of language material.

Basic data	**HK-P**	**PUD**	**PUD-N**	**PUD-W**	**GSD**	**LCMC**
Number of sentences	354	1,000	500	500	3,997	45,590
Number of word tokens*	4,303	17,844	8,699	9,145	80,978	827,625
Number of word types (in Chinese characters)*	778	4,943	2,876	3,081	15,815	42,506

*excluding punctuation marks and words including non-Chinese graphemes (e.g. Latin letters, Arabic numerals, symbols)

3.2 Language units

The chapter describes the determination and operationalisation of language units we chose to analyse.

3.2.1 The sentence

The sentence is represented in UD as a tree, which is built on asymmetric and directed binary relations represented by tree edges between words represented by tree nodes (e.g. Nivre et al., 2020, p. 4035; de Marneffe et al., 2021, p. 257). Only one word is promoted to be a head

[19] We decided not to analyse the fourth UD Chinese CLF treebank (Lee, Leung and Li, 2017) because it includes essays written by non-native speakers learning Chinese.

of the whole sentence – called root – while the rest of the words directly or indirectly – through other words – depends on it.

3.2.2 The clause

The simple clause (de Marneffe et al., 2021, pp. 272-276) consists of a head, i.e. verbal or non-verbal predicate, and its directly or indirectly dependent words (if any). The simple clause can correspond to a sentence with only one predicate (a root) and, consequently, can be represented by a whole tree. Otherwise, it is a subtree corresponding to the main clause or a clause integrated into a sentential structure through coordination or subordination. The determination of coordinate or subordinate clauses relies on the UD annotation for particular dependency relations that their predicates carry. In the case of coordination, if a predicate governs a word which depends on it via the UD conjunct relation (`conj`), we consider the dependent word to be a predicate of another – coordinate – clause. When it comes to subordination (de Marneffe et al., 2021, pp. 277-280), UD distinguishes five basic relations assigned to a predicate of a subordinate clause – clausal subject (`csubj`), clausal complement (`ccomp`, `xcomp`), adverbial clause modifier (`advcl`) and adnominal clause modifier (`acl`). In addition, the clausal syntactic relation can also occur in a special form of parataxis.

It should be noted that we utilise an approach to the clause which disregards the dependency relation between clauses, or more precisely, the edge between a head of a subordinate clause and its governor. Consequently, it treats each clause separately. This exclusive approach (applied by Köhler and Naumann, 2009; Berdicevskis, 2021; or used in Prague Dependency Treebank 3.0, Bejček et al., 2013) prevents multiple processing of the same sentential segments.

3.2.3 The syntactic phrase

In general, the syntactic phrase (or shortly phrase) represents any subtree starting with a word (a node) being a phrasal head and continuing with other – directly or indirectly – dependent words (nodes). Regarding its determination, we follow an approach introduced by Mačutek, Čech and Milička (2017). The authors determined the phrase as a complete subtree directly hanging from a predicate of the main clause, while predicates of coordinate or subordinate clauses were disregarded due to annotation limits of analysed language material. Since we can distinguish coordinate or subordinate clauses in the UD treebanks, we approach the syntactic phrase in two different ways.[20] Firstly, we precisely follow Mačutek, Čech and

[20] We are fully aware that the coordination concerns not only with the clausal but also phrasal level. When determining a clause, we rely on UD annotation for dependency relations. As regards the determination of a coordinate clause, we use the UD dependency relation of the conjunct (`conj`), (cf. Berdicevskis, 2021). However, when determining the phrase, we rely on structures of dependency trees. Except for the need to identify the coordinate clause, we do not aim to investigate the relation between the coordination and

Milička (2017), i.e. only phrases directly depending on a head of a sentence (i.e. a root) are taken into account. The syntactic phrase is viewed as a complete subtree in this case – starting with its phrasal head and ending with its terminal node(s). Due to the fact that it directly hangs from the root of a sentence, we term it a 'sentential' phrase. Secondly, we apply the same approach to all clausal heads identified within a sentence (Berdicevskis, 2021). Hence, this approach treats the phrase as a subtree that hangs from the head of each simple clause. The phrase cannot be the clause itself,[21] and any clause embedded into it is excluded. Both conditions prevent multiple processing of the same sentential segment which would act as a phrase or its integral part and then as the clause itself. We term the phrase 'clausal'.

Finally, we also follow an alternative approach proposed by Mačutek, Čech and Courtin (2021), who determine a unit corresponding to the phrase level as "the longest possible sequence of words (belonging to the same clause) in which all linear neighbours (i.e. words adjacent in a sentence) are also syntactic neighbours (i.e. they are connected by an edge in the syntactic dependency tree which represents the sentence)" (Mačutek, Čech and Courtin, 2021, p. 3).[22] The authors term the unit as a linear dependency segment (LDS).

3.2.4 The word

UD and its annotation build on dependency relations between words (de Marneffe et al., 2021, p. 257), representing nodes in a dependency tree and carrying morphosyntactic annotation. Generally, the Chinese word in the UD and LCMC samples corresponds to a string of Chinese – traditional or simplified – characters.

3.2.5 The character, component and stroke

The Chinese character represents a basic graphic unit of the Chinese script and corresponds to a syllable with one exception (see Chapter 3.2.6). Its structure is divisible either into components or strokes. The inventory of strokes for each character is immutable, whereas the inventory of components depends on a chosen segmentation strategy. To process the character length, we decided to use an open-source document published by Beijing Language and Culture University which contains a list of components and the number of strokes for each

the law and test the impact of the coordination on the results. It is another complex theoretical issue which can be approached in several ways (cf. Osborne, 2019), hence, we do not go into the depth and take the coordination into account on the phrasal level.

[21] C.f. "phrases are distinguished from clauses mainly by the absence/presence of a finite verb" (Osborne, 2019, p. 6).

[22] Due to the word order that the approach takes into account, the linear dependency segment does not entirely correspond to the phrases mentioned above, which determination relies on the syntactic dependency criterion. However, due to its position in the unit hierarchy corresponding to a level between the clause and word, we include the linear dependency segment into chapters on the syntactic phrase.

of 6,647 Chinese characters (shortly BLCU).[23] However, to use the document, all words in the samples must be written in simplified Chinese characters. While words in UD Chinese GSDSimp treebank and LCMC consist of simplified characters, UD Chinese HK and PUD treebanks were converted into their simplified forms by virtue of available software (文林 Wénlín Software for Learning Chinese: Version 4.0.2, 2011).

3.2.6 The syllable and sound

The Chinese syllable consists either of a vowel or a combination of a vowel, glide(s) and/or consonant(s) (Wee and Li, 2015, p. 475). It corresponds to a Chinese character with one exception, i.e. erization, which is captured by one syllable but two characters, e.g. 这儿 *zhèr* 'here'). Due to this high correspondence and the fact that to determine the number of syllables in a word (not syllable boundaries) is sufficient from the menzerathian perspective, the Chinese characters, which are primarily used to capture Chinese words, can be the only measurement unit of the word (applied, for example, by Chen and Liu, 2016). As far as erization is concerned, we disregard quantitative differences between characters and syllables because erization occurs in our samples to a minimal extent.[24]

The determination of the sound relies on the International Phonetic Alphabet (IPA). We firstly automatically converted the Chinese characters into pinyin, i.e. Hanyu Pinyin, 'Chinese Phonetic Writing', by virtue of an open-source tool, a Python library pypinyin (Python-pinyin, 2022). Secondly, we compared both the alphabetic systems to identify those cases when one pinyin letter does not correspond to a sound in IPA, or in other words, there is no one-to-one correspondence between them. Based on the identified differences, we drew up several rules (Lin, 2007, pp. 121-129) for developing an algorithm which automatically alters pinyin, i.e. uses an alternative symbol to lengthen or shorten the pinyin transcription (see Table 3).

[23] 汉字信息词典 (Dictionary of Chinese Character Information), accessed: December 2, 2021.

[24] HK-P does not contain any case of erization. PUD contains four cases out of 17,844 word tokens, GSD 16 cases out of 80,978 word tokens and LCMC 663 cases out of 827,625 word tokens.

Table 3. Overview of quantitative differences between pinyin letters and sounds in IPA.

Sound type	Pinyin	IPA	Number of letters	Number of sounds	Difference *
Post-alveolar affricate	ch, zh	tʂʰ, tʂ	2	1	-1
Post-alveolar fricative	sh	ʂ	2	1	-1
Velar nasal	ng	ŋ	2	1	-1
Labial consonant b, p, m, f + vowel	o	wo	1	2	+1
Diphthong	ai, ao, ei, ou	ai̯, ɑu̯, ei̯, ou̯	2	1	-1
Consonant + vowel + velar nasal ng	i	jə	1	2	+1
Consonant + vowel + alveolar nasal n	u	wə	1	2	+1
Glide + e/an	yu	ɥ	2	1	-1

*when IPA is compared to pinyin

3.3 Language unit combinations and their quantification

The following section introduces the measurement units, i.e. direct constituents, which we opt for all the constructs.

Sentence

Measurement unit: clause – The sentence length is measured in the number of clausal heads, i.e. words which carry the dependency relations of `root`, `csubj`, `ccomp`, `xcomp`, `acl`, `advcl`, `parataxis` or `conj` if it inherits the predicate function.

Measurement unit: sentential phrase – The length of the sentence is expressed as the number of nodes which directly depends on a root of a sentence. Sentences consisting only of the root are disregarded because their lengths equal zero. The root is not considered to be the phrase.

Clause

Measurement unit: word – The clausal length is calculated as a sum of words a) which directly or indirectly (through other words) depend on a clausal head and b) which do not belong to another clause. The clausal head is included in the sum of words in the clause.

Measurement unit: clausal phrase – In this case, we count all words a) which directly depend on the clausal heads (`root`, `csubj`, `ccomp`, `xcomp`, `acl`, `advcl`, `parataxis` or conj with the predicate function) and b) which are not the clausal heads themselves, i.e. do not carry these clausal dependency relations. The clausal head is not determined as the phrase.

Measurement unit: linear dependency segment (LDS) – The length of the clause is expressed as the number of LDSs identified as the longest possible chains of words which are connected syntactically in a dependency tree (i.e. by an edge) while respecting the word order in the clause. LDS includes the clausal head.

Syntactic phrase

Measurement unit to the sentential phrase: word – The length of the phrase is expressed as a sum of words which includes a word directly dependent on the root (i.e. a phrasal head) and all other words directly or indirectly (through other words) dependent on it.

Measurement unit to the clausal phrase: word – This phrase is also measured as a sum of the words. However, the sum includes 1) a node which directly depends not only on the root but also on other clausal heads (csubj, ccomp, xcomp, acl, advcl, parataxis or conj with the predicate function) and 2) words which are directly or indirectly (through other words) dependent on it unless they belong to another clause.

Measurement unit to the linear dependency segment (LDS): word – The length of LDS is expressed as the number of words which are connected via dependency relations and are linear neighbours. Even though the punctuation marks are included in dependency trees as integral nodes, they do not interrupt the dependency relations or linear neighbourhood between the words.

Word

Measurement units: character/syllable – The word length is always measured as the number of Chinese characters corresponding to syllables in Chinese only except for erization.

Character

Measurement unit: component or stroke – The length of the simplified Chinese character is calculated either as a sum of its components (based on the BLCU source), each of which consists of a partial number of strokes, or as a total number of all strokes.

Syllable

Measurement unit: sound – The syllable length is expressed as a sequence of letters and/or symbols representing sounds in IPA.

3.4 Testing the model reliability

Based on the quantification of the language material, the construct length, its frequency, and the mean constituent length are calculated. To avoid possible biased results by these so-called outliers, we treat them with the method of the weighted average (e.g. applied by Mačutek, Čech and Courtin, 2021). If the frequency of a construct length is lower than 10, we pool the construct with its shorter neighbour(s) until their frequency sum meets our requirement (i.e.

being equal or greater than 10). The lengths of the construct and constituent are subsequently calculated as the weighted average of the pooled values while using the frequency as their weights. We fit the weighted values with two models proposed by Altmann (1980), i.e. the complete model $y(x) = ax^b e^{cx}$ with three parameters *a*, *b*, and *c,* and the truncated model $y(x) = ax^b$ with the parameter a being replaced by the constituent length of the one-constituent construct y_1 (c.f. Kelih, 2010; Čech and Mačutek, 2021), and the parameter *b*. The NLREG Version 6.3 (Sherrod, 2005) software is used for the fitting of both the mathematical models to data in order to obtain values of the parameters and the coefficient of determination R^2. We interpret the goodness-of-fit as reliable if the coefficient of determination R^2 reaches the value equal to or greater than 0.90 (Mačutek and Wimmer, 2013, p. 233).

4 Menzerath-Altmann law applied

The chapter brings results which we yield for chosen unit combination. Scripts created for data processing and all processed data (including their non-weighted versions) are available on Github.[25]

4.1 The sentence as the construct

Hypotheses:

1. The longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in words.
2. The longer the sentence length measured in the number of sentential phrases, the shorter the mean length of the sentential phrases measured in words.
3. The longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in clausal phrases.
4. The longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in linear dependency segments (LDS).

The results of each triplet on the sentence level corroborate the hypotheses and the coefficients of determination R^2 meet standard of $R^2 \geq 0.90$ with only two exceptions.

Despite the hypothesis's corroboration, the triplets differ in evaluating the construct and constituent lengths based on the limits of the short-term memory span (7 ± 2, Miller, 1956). When opting for the clause as the direct measurement unit for the sentence, the GSD sample suffers from the wide scale of sentence lengths which considerably exceed the upper threshold of the short-term memory span. This issue does not arise when the sentence is measured directly in sentential phrases. However, the mean lengths of the sentential phrases measured in words exceed this upper limit themselves. The triplet of the sentence, clause and word struggles with the same issue – the mean clause lengths are too long, which puts the granularity of both the triplets into question. The phrase does not appear to be the direct measurement unit for the sentence and the word for the clause.

Using the clausal phrase and the linear dependency segment as the direct measurement units of the clause sufficiently lowers the mean clause lengths to meet the limits of the short-term memory span. Hence, the sentence, clause and phrasal unit appear to be the appropriate unit combination. On the one hand, both the triplets – sentence, clause and clausal phrase / linear dependency segment – still face the wide scale of the sentence lengths in GSD. On the other hand, this wide scale might be caused by a different factor (or factors) coming into play. For example, the alternative determination of the clause based on selected punctuation marks solves this issue while still corroborating the hypothesis. These results indicate the specificity of the UD annotation of the clausal dependency relations.

[25] Available at https://github.com/TerezaMotalova/menzerath-altmann_law_in_chinese.

When comparing the clausal phrase and linear dependency segment, we cannot unambiguously conclude based on the goodness-of-fit which unit achieves better results. However, if we compare their determinations, the clausal phrase faces the issue of disregarding clausal heads – they are neither parts of the phrases nor the phrases themselves, whereas the linear dependency segment does not leave any word out of the analysis. Nevertheless, the clause and the phrase (i.e. clausal phrase and linear dependency segment) have to be further tested to shed light on their behaviour when their positions in the unit hierarchy change.

The question also arises why the goodness-of-fit is above the standard (i.e. $R^2 \geq 0.90$) for the triplet of the sentence, clause and word as well as the triplet of the sentence, clause and phrase (either clausal phrase or linear dependency segment) when their sub-constituents, i.e. the word and the phrase, are not obviously of the same level. The hypothesis's corroboration for both the triplets leads to an assumption that skipping a level in the case of a sub-constituent does not always have a considerable impact on the results.

4.2 The clause as the construct

Hypotheses:

1. The longer the clause length measured in the number of words, the shorter the mean length of the words measured in (Chinese) characters.
2. The longer the clause length measured in the number of clausal phrases, the shorter the mean length of the phrases measured in words.
3. The longer the clause length measured in the number of linear dependency segments (LDS), the shorter the mean length of LDSs measured in words.

Going one level below in the vertical hierarchy of the language units brings opposite results in comparison with the sentence level. The goodness-of-fit between the models and the data is unsatisfactory, and the hypothesis is rejected in most cases when the clause becomes the construct.

In the case of the triplet of the clause, word and (Chinese) character, the clause lengths suffer from the wide scale which extensively exceeds the upper threshold of short-term memory span (i.e. 7 ± 2, Miller, 1956). This supports the assumption made on the sentence level that the word is not the direct constituent of the clause. On the contrary, the mean word lengths suffer from the narrow range of one to two Chinese characters, which reflects the word length distribution in Chinese and poses a question of whether the prevalence of these words represents the boundary condition for the law.

As for the triplets including the clausal phrase and linear dependency segment, both the approaches show their pros and cons on this level. When it comes to the former, on the one hand, the clause lengths do not exceed the upper limit of the short-term memory span. On the other hand, the determination of the clausal phrase leads to the exclusion of words functioning as clausal heads because they are neither part of the phrases nor the phrases themselves. Including clausal heads with at least one phrase into mean phrase lengths demonstrates that the determination seriously impacts the results. The mean phrase lengths start decreasing after the heads are included. Nevertheless, the inclusive approach faces methodological drawbacks and is only illustrative.

The determination of the linear dependency segment does not leave any word out of analysis but struggles with clause lengths crossing the upper threshold of the short-term memory span. LDSs are determined based not only on the dependency syntactic criterion but also on the criterion of the linear neighbourhood. Hence, clauses are more fragmented and consist of a higher number of constituents compared to the clausal phrases.

When comparing both the approaches with respect to the coefficient of determination R^2, the triplet including the linear dependency segment mostly yields better results. However, if we consider the alternative approach to the clausal phrase, which does not disregard the clausal heads with at least one phrase, most of the coefficients of determination R^2 reach higher values than in the case of the linear dependency segment. To sum it up, at least one unit exists between the clause and the word – the phrase. However, its determination faces several issues to tackle.

4.3 The phrase as the construct

Hypothesis:

1. The longer the length of a phrasal unit measured in the number of words, the shorter the mean length of the words measured in (Chinese) characters.

The chapter presents results obtained by analysing three different units being the construct on the phrasal level – sentential phrase, clausal phrase and linear dependency segment. When we test the sentential phrase measured in words, the poor results and excessively long lengths being above the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) indicate that the phrase as a subtree directly dependent on a root is not the direct constituent of the sentence and a linguistic level is skipped (e.g. a clause).

As for the clausal phrase and the linear dependency segment, the mean word lengths start to decrease only after the frequency of unit usage is disregarded, or in other words, types are analysed. Moreover, the triplet including the linear dependency segment corroborates the hypothesis in most of the samples. The results clearly show that mean word lengths are able to decrease in the menzerathian trend despite the prevalence of one- and two-character words in our samples and generally in Chinese, which initially appeared to be the boundary condition for the law. Therefore, the unit frequency is the decisive factor in whether the Menzerath-Altmann law comes into force after all. In addition, LDS represents the first unit on the phrasal level whose lengths respect the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956).

Finally, the homogeneity of the samples represents another important factor for the law when the types are analysed. Excluding phrases containing at least one non-Chinese grapheme improves the results that are not achieved when the law is applied to all phrase types.

4.4 The word as the construct

Hypotheses:

1. The longer the word length measured in the number of Chinese characters, the shorter the mean length of the characters measured in components.[26]
2. The longer the word length measured in the number of Chinese characters, the shorter the mean length of the characters measured in strokes.
3. The longer the word length measured in the number of syllables, the shorter the mean length of the syllables measured in sounds.

The chapter presents the results of the word in the position of the construct. Its length is always measured in Chinese characters roughly corresponding to syllables, while its sub-constituent changes to the component, stroke and sound. The results of the triplets show that the law is firstly highly sensitive to word segmentation which disables or enables the law to reveal its behaviour. Secondly, the law manifest itself or the menzerathian decreasing tendency appears when only word types are analysed. Or in other words, the law is sensitive to the frequency of unit usage. In the case of the tokens, mean character (syllabic) lengths of one-character (syllable) words have lower or even the lowest values, or the overall trend is increasing. On the one hand, such results accord with the Brevity law preferring the usage of shorter units. On the other hand, they contradict the Menzerath-Altmann law. Based on the results of the types, we can conclude that the prevalence of one- and two-character words in Chinese does not represent a boundary condition for the Menzerath-Altmann law, even if the word is the construct measured directly in Chinese characters (cf. Chen and Liu, 2022).

When comparing the results of the types, the UD samples yield unsatisfactory results for the triplets including the component and stroke but corroborate the hypothesis at least by one model if the triplet includes the sound. The LCMC samples show the opposite. While they do not reject the hypothesis for the component and stroke, they mostly do for the sound. Differences in the results of the types also indicate that other factors influence the law. Firstly, when considering decomposing approaches to Chinese characters, the best fitting results are achieved when characters are maximally decomposed (i.e. until each component cannot be decomposed further). Secondly, an LCMC sample containing texts only of one text type always corroborates the hypothesis, while mixed LCMC samples do not. These results indicate that sample homogeneity (or heterogeneity) is another factor coming into play. Thirdly, mean character lengths of words having seven and more characters deviate from the decreasing trend. The question arises whether we face an issue of compound words which behave irregularly with regard to the law (Mačutek and Rovenchak, 2011).

Finally, mean character lengths of the word types show an apparent decreasing trend regardless of whether they are measured in components or strokes. These results contradict the assumption that skipping a level leads to an increase in constituent lengths or at least their irregular behaviour. Leaving a linguistic level out might not always have a significant impact

[26] We exclude all words containing at least one non-Chinese grapheme from the analysis, which applies to all triplets and all samples tested on the word level.

when it comes to a sub-constituent (cf. the sentence level, Chapter 4.1). On the other hand, one-character words are expected to be composed of characters having the highest number of components on average. If we add up their number of strokes, the sums would be the highest, or in other words, these words would be composed of characters having the highest number of strokes on average. A graphic field in which a character must fit exerts strong pressure on the character due to its limited size. Hence, the character must sufficiently self-regulate and self-organise itself to ensure its readability. While the number of components can change within a character, the number of strokes cannot. Hence, there is a simple principle – the more components a character has, the lesser stroke the components have. From this perspective, both the units appear to be on the same level in the hierarchy of language units. Only scales of their lengths differ and the stroke might be a more stabilised unit.

4.5 The character as the construct

Hypothesis:

1. The longer the Chinese character length measured in the number of components, the shorter the mean length of the components measured in strokes.[27]

The triplet consisting of the character, component and stroke is the only unit combination which corroborates the hypothesis for both – the tokens and the types from the UD samples and LCMC. The goodness-of-fit between the models and all the data meets the standard of $R^2 \geq 0.90$. Moreover, the character tokens and types corroborate the hypothesis without regard to a decomposing approach, i.e. decomposition based on BLCU source and maximal decomposition (each character into its components until all identified components could not be decomposed further). When it comes to short-term memory, neither $ChLs$ nor $MCoLs$ exceed its upper threshold (i.e. 7 ± 2, Miller, 1956), which applies to both – the character tokens and types – and all the samples under analysis on this level.

Finally, the corroboration of the hypothesis by the tokens poses a question of why the Brevity law does not come into force. Compared to higher linguistic levels, the Chinese character is a basic graphic unit of the Chinese script which is organised within a graphic field of limited size. The reverse tendency – the higher the number of components, the higher the number of strokes on average – cannot apply because the character needs to fit in the graphic field while being readable and distinguishable from other characters. If most characters follow such self-regulation and self-organisation, the frequency of usage – the Brevity law – does not prevent the Menzerath-Altmann law from coming into force.

[27] All non-Chinese graphemes are excluded from the analysis.

4.6 The parameters $\boldsymbol{a}$ and $\boldsymbol{b}$

This last chapter presents the results of the parameters a and b of the truncated formula, which we yielded throughout the whole hierarchy of analysed language units, i.e. sentence, clause, phrase, word, character/syllable, component and stroke (see Figure 1). Based on these results, several conclusions can be drawn. However, it is important to emphasise that the conclusions are only preliminary due to issues which arose in relation to the determination and neighbourhood of language units belonging to particular unit triplets.

Values of both the parameters appear to be, first and foremost, under the influence of a linguistic level or even levels involved in a unit triplet. To illustrate the point, we can take the word level as an example. Using the component and sound as the measurement units for the character/syllable keeps its values clustered together, whereas opting for a stroke results in their increase. As regards the influence on the parameter b, higher linguistic levels tend to yield lower values (e.g. sentence vs word). The parameter also seems to be determined by variability in constituent lengths. Its lowest values are observed on the sentence level, where the clause and the phrase occupy the position of the direct constituent. Measuring both in words leads to a higher variance in their lengths and a steeper decrease. On the contrary, variability in constituent lengths of the word, i.e. the character/syllable measured in components/sounds, is lower and the lengths decrease gradually. The parameter b has the highest values in this case. However, not only the linguistic level but also its determination comes into play. To illustrate the point, we can take the sentence measured in clauses as an example. When the mean lengths of clauses are measured in clausal phrases, the parameter a reaches lower values and parameters b reaches higher values. Mean clausal lengths measured in linear dependency segments show the opposite – higher values of the parameter a and lower values of the parameter $b.$ The results also reveal that values of the parameters from lower linguistic levels (e.g. word or character) more or less cluster together. In comparison, values from higher linguistic levels (e.g. sentence) are dispersed to a greater degree. Hence, lower levels appear to be more stabilised in a language system (e.g. the word), whereas higher levels show a higher degree of variability (e.g. sentence). The variability in lengths might enable other factors to come into play or amplify their impact on the results, for example, a text.

As regards the relationship between the parameters, their values tend to be negatively correlated – not only within linguistic levels separately but also across the levels. If we apply the Kendall rank correlation test to all values of the parameters (variables are not normally distributed), a value of Kendall's τ coefficient equals -0.56 while the *p*-value < 0.001. The correlation is statistically significant and can be classified as a moderate negative correlation, i.e. -0.50 to -0.70 (Hinkle, Wiersma and Jurs, 2003).

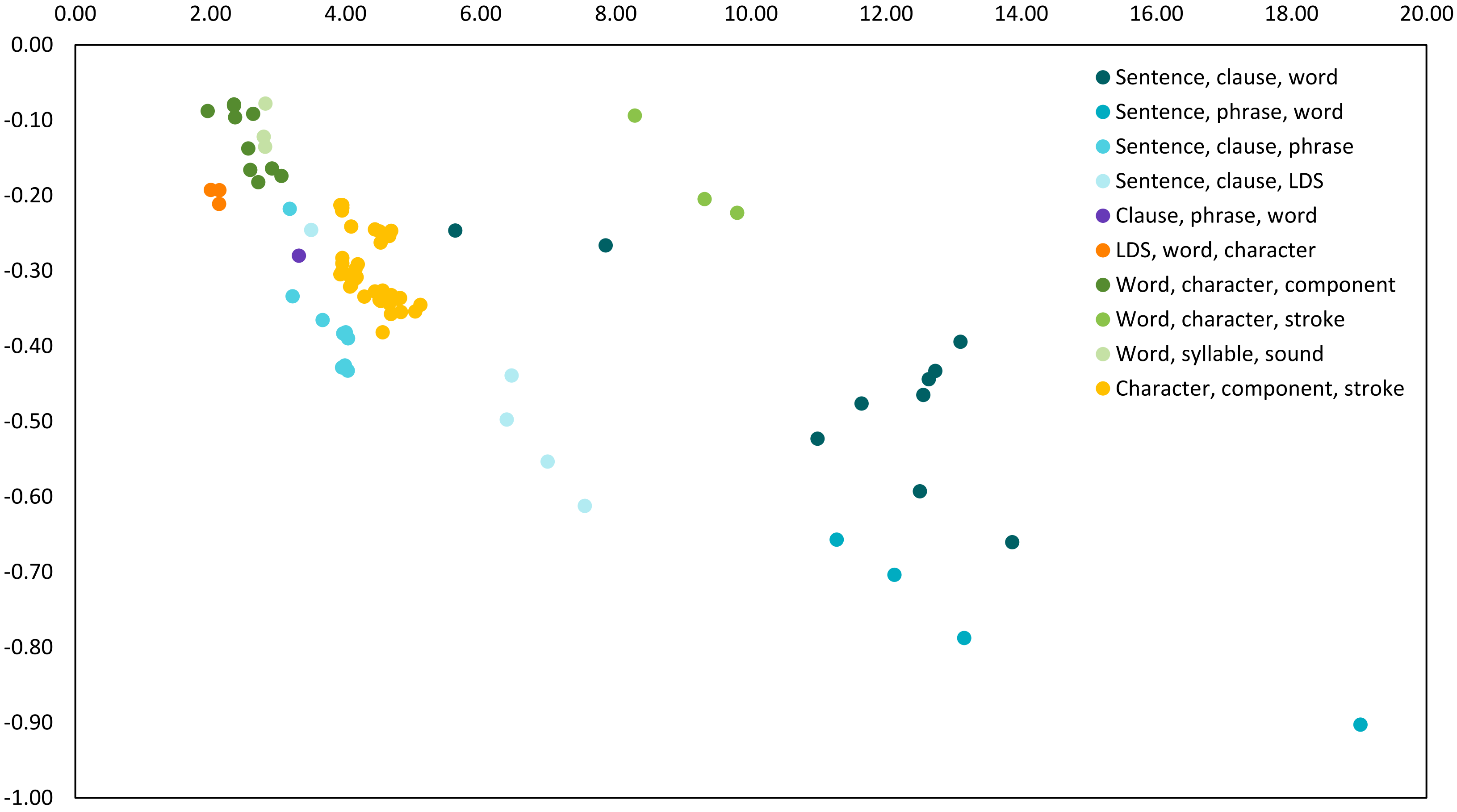

Figure 1. Visualisation of the parameters a and b of the truncated model obtained from all linguistic levels under analysis.

Conclusion

The thesis focused on the application of the Menzerath-Altmann law according to which lengths of two language units of different hierarchical levels – a hierarchical higher construct and a hierarchical lower constituent – are negatively correlated. The thesis applied the law to Chinese and pursued general and language-specific objectives. First, a hierarchy of language units, i.e. sentence, clause, phrase, word, character/syllable, component/sound and stroke, was tested to observe how the units which are not peripheral behave when they switch their hierarchical position from the constituent to the construct. Second, it is generally assumed that the negative correlation between lengths of two language units appears as far as immediately neighbouring units are involved. Or in other words, a linguistic level between them is not skipped. However, it is not always unambiguous whether two language units can be considered immediate hierarchical neighbours. Hence, the second objective was to test various unit combinations to shed light on the unit neighbourhood. Thirdly, considering that the law is a general mechanism maintaining equilibrium in cognitive workload, we also evaluated construct and constituent lengths based on Miller's 'magical number plus or minus two' (1956), representing the maximum amount of information which we can process in short-term memory. Fourthly, the clause and the word are preferred to be immediate hierarchical neighbours in studies on Chinese. Hence, the fourth objective of the thesis was to include the phrase level (determined as sentential phrase, clausal phrase and linear dependency segment, shortly LDS) into the hierarchy of language units in Chinese and test its behaviour towards other units when its hierarchical positions change. Finally, Chen and Liu (2016, 2019, 2022) yielded that the law does not come into force when the word and the Chinese character are tested as the construct and the constituent accordingly. Based on the results, the prevalence of one- and two-character words in Chinese appears to be a boundary condition for the law to manifest itself. Hence, the last objective was to examine whether other factors (e.g. frequency) prevent the law from coming into play.

Based on the results which we yielded by testing the law throughout the whole hierarchy of language units mentioned above, we have come to the following conclusions:

- As regards the behaviour of non-peripheral language units with regard to their different hierarchical positions, the results showed that the law can be corroborated for a given language construct and its constituent but rejected when the constituent switches its hierarchical position over to the construct. All unit combinations on the sentence level corroborated the law (i.e. sentence, clause, word; sentence, phrase, word; sentence, clause, phrase/LDS). However, the clausal level yielded opposite results. The law was rejected when the clause measured in words/clausal phrases/LDS and the sentential phrase measured in words became the constructs. The trend in the results can also be reverse. While the combination of the clause, LDS and word did not corroborate the law, LDS becoming the construct and the word and character becoming its direct and indirect constituents mostly did. All these contradictory results across the levels amplify the need to test a given language unit in its different hierarchical positions.

- When it comes to the unit neighbourhood, the achieved results revealed that the sentence and phrase do not appear to be immediate hierarchical neighbours as well as the clause and word. On the one hand, each unit combination on the sentence level corroborated the law. On the other hand, constituents of the sentence differed in their lengths when being evaluated based on the upper threshold of the short-term memory span, i.e. Miller's 7 ± 2 (1956). While the mean lengths of the clause and the phrase both measured in words exceeded the upper threshold, the mean clause lengths measured in phrases or LDSs were in accord with it. These results indicated that the phrase might not be an immediate hierarchical neighbour for the sentence and the word for the clause. This assumption was supported when the clause and phrase measured in words became the constructs. Their lengths excessively exceeded the upper threshold and the law was rejected. Although the law was also rejected for the clause measured in clausal phrases, its lengths respected the short-term memory limit and indicated that at least one unit exists between the clause and the word – the phrase. However, its determination faces several issues to tackle (see below). To sum it up, Miller's 'magical number plus or minus two' might be considered a rule of thumb for evaluating the construct and constituent lengths. Agreement with this limit might indicate the neighbourhood and/or an appropriate determination of a chosen unit, especially for higher linguistic levels (cf. Jiang and Ma, 2020; Mačutek, Čech and Courtin, 2021).

- The determination of a language unit represents another important factor for the law. Let us start with the clausal phrase and linear dependency segment. The former was determined based on the dependency syntax as a sum of all words that (directly or indirectly) depend on a clausal head unless they belong to another clause (Mačutek, Čech and Milička, 2017). The length of the latter was expressed as a sum of words which are connected through dependency relations and are linear neighbours in a clause (Mačutek, Čech and Courtin, 2021). Both the phrasal units were tested in three different positions within the following combinations – 1) sentence, clause, phrase; 2) clause, phrase, word; and 3) phrase, word, Chinese character. In the case of the sentence level, the law was corroborated and the impact of the phrase determination appeared to be minimal. On the contrary, the law was rejected on the clause level where both the approaches revealed their pros and cons. In the case of the clausal phrase, on the one hand, clause lengths did not exceed the upper limit of short-term memory (7 ± 2, Miller, 1956). On the other hand, the determination excluded words functioning as clausal heads from the analysis because they were neither part of the phrases nor the phrases themselves. The linear dependency segment showed the opposite – its determination did not leave any word out, but clause lengths crossed the upper threshold of short-term memory. Finally, in the case of the phrase level where both the units were in the construct position, the law started to manifest itself. Or in other words, their mean word lengths started to decrease. However, only if the frequency of unit usage was disregarded, or in other words, the phrase types were analysed. Moreover, the linear dependency segment was the only unit on the phrasal level that corroborated the law in most cases and whose lengths followed the upper threshold of short-term memory. Despite the drawbacks, we can preliminarily conclude that the phrase can be a legitimate unit in the unit hierarchy in Chinese and that the prevalence of one- and two-character Chinese words does not prevent the law from coming into force when the word is in the constituent position.

- The sensitivity of the law to the unit determination also appeared on the word level. We tested the word in the construct position on two sets of samples. The first included Universal Dependencies treebanks (Zeman et al., 2021). Samples of the second set came from the Lancaster Corpus of Mandarin Chinese (McEnery, Xiao and Mo, 2003). Both the sources implied different approaches to word segmentation and yielded contradictory results when the law was applied to the unit combination of the word, character and component/stroke. While the set of samples from the Universal Dependencies rejected the law, the set of samples from the Lancaster Corpus of Mandarin Chinese did not. The results indicated that the word segmentation represented a crucial factor which disables or enables the law to manifest itself. The impact of the word segmentation also appeared in connection with words whose lengths were equal to or greater than seven and more characters. Their compound forms apparently caused deviation of their mean character lengths from the menzerathian decreasing trend (cf. Mačutek and Rovenchak. 2011). Finally, the results on the word level showed that different approaches to the decomposition of Chinese characters into their components influence the degree of agreement between empirically obtained results and theoretical results predicted by the law. The law was always corroborated when the Chinese characters were maximally decomposed (until the components of each character could not be decomposed further).

- Not only phrases but also words being constructs showed that the law manifested itself or the menzerathian decreasing tendency appeared when the frequency of unit usage was not taken into account, in other words, only when types were analysed. When the law was applied to phrase and word tokens, constituents belonging to the shortest constructs had lower values than the following constituent lengths, which contradicted the law. The analysis of unit tokens reflects the competition between the Menzerath-Altmann law and the Brevity law. While the former law expects constituents of the shortest construct to be the longest, the latter law predicts the negative correlation between the unit length and its frequency. Hence, constituent lengths can be lowered by shorter units which are more frequent. The analysis of the word types showed that the prevalence of one- and two-character words in Chinese does not represent the boundary condition for the Menzerath-Altmann law, even if the word is in the position of the construct. Based on these results, we can also conclude that the Chinese character can be regarded as an immediate hierarchical neighbour of the word (cf. Chen and Liu, 2022, who left the Chinese character out of the hierarchy and measured the word tokens in components).

- The sample homogeneity can also be another decisive factor for the law, as demonstrated on the phrase level. When the word measured in characters became the constituent, the issue of words fully or partly consisting of non-Chinese graphemes arose. While one Chinese grapheme, i.e. Chinese character, roughly corresponds to a syllable, one non-Chinese grapheme usually represents a letter, numeral, or symbol. Applying the law to phrase types consisting solely of Chinese characters considerably improved the agreement between empirical and theoretical results compared to the agreement yielded by testing all phrase types.

- The so-called truncated model of the law includes two parameters – the parameter a (the mean constituent length of the shortest construct) and the parameter b. It has been shown that their values tend to be negatively correlated (e.g. Hou et al., 2019a; Jiang and Jiang, 2022). Hence, we used values of both the parameters obtained from all unit combinations that corroborated the law, and statistically tested their relationship (by the Kendall rank correlation test). The results showed that the correlation is statistically significant and can be classified as a moderate negative correlation (Hinkle, Wiersma and Jurs, 2003).

Finally, if we were to draw only one conclusion about the results presented in the thesis, then Menzerath-Altmann is not only about its application to any language material but, first and foremost, about considering competitive and cooperative factors which might have an impact on the results and cast light on the behaviour of language units under analysis and the law itself.

Annotation in the Czech language

Disertační práce se věnuje Menzerathovu-Altmannovu zákonu a jeho aplikaci na čínský jazykový materiál. Tento zákon předpokládá, že délky dvou jazykových jednotek, tj. hierarchicky vyššího konstruktu a hierarchicky nižšího konstituentu, spolu negativně korelují. Během posledních čtyř desetiletí byla platnost zákona ověřena na různých jazykových jednotkách a různém jazykovém materiálu. Určité jednotky však byly testovány častěji než jiné a většinou pouze v jedné hierarchické pozici, tj. konstruktu nebo konstituentu. Negativní korelace predikovaná Menzerathovým-Altmannovým zákonem by se měla objevovat mezi délkami jednotek bezprostředně sousedících jazykových rovin. Vymezení jednotlivých rovin a jejich jednotek ale není vždy zřejmé a jednoznačné. Cílem této disertační práce je testování Menzerathova-Altmannova zákona napříč hierarchií jazykových jednotek, která je složená z věty, klauze, fráze, slova, znaku/slabiky, komponentu/hlásky a tahu. Práce nejprve analyzuje chování jednotek v obou hierarchických pozicích, tj. konstituentu i konstruktu (s výjimkou věty, slabiky, komponentu a tahu). Zároveň jsou analyzovány různé kombinace jednotek z důvodu náležitého vymezení hranic mezi jednotlivými rovinami a jejich jednotkami v čínštině. V neposlední řádě práce zkoumá faktory (např. frekvenci), které omezují platnost zkoumaného zákona. Provedené analýzy přinesly několik závěrů. Délky jednotek na rovinách věty, fráze, slova a čínského znaku v pozici konstruktu se v provedených analýzách chovají v souladu s Menzerathovým-Altmannovým zákonem. V případě klauze v pozici konstruktu se ukázalo problematickým vymezení jejích bezprostředních jednotek. Konkrétně, z perspektivy Menzerathova-Altmannova zákona se slovo neprojevuje jako bezprostředně sousedící jednotka klauze. Předběžně lze na základě provedených analýz konstatovat, že touto bezprostřední jednotkou je syntaktická fráze. Analýzy dále ukazují, že několik zásadních faktorů ovlivňuje platnost daného zákona. Zaprvé, v případě frází a slov se zákon projevuje pouze tehdy, pokud se analyzují tzv. typy (types) a nikoli tokeny (tokens), tj. nebere se v úvahu frekvence. Zadruhé, platnost zákona závisí na způsobu segmentace na slova (tzv. tokenizace) analyzovaného jazykového materiálu. Zatřetí, zákon je citlivý na homogenitu jazykového materiálu. Při testování Menzerathova-Altmannova zákona je tedy důležité zohlednit konkurující a kooperující faktory, které mohou jednak ovlivnit výsledky, jednak poodhalit chování jazykových jednotek i samotného zákona. Analýzy byly provedeny na dependenčně syntakticky anotovaném jazykovém materiálu.

Bibliography

Altmann, G. (1980) "Prolegomena to Menzerath's law", in Grotjahn, R. (ed.) *Glottometrika 2*, Studienverlag Dr. N. Brockmeyer, pp. 1-10.

Altmann, G. (1983) "H. Arens' 'Verborgene Ordnung' und das Menzerathsche Gesetz", in Faust, M. et al. (eds.) *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik*, Gunter Narr Verlag, Tübingen, pp. 31-39.

Altmann, G. (1992) "Das Problem der Datenhomogenität", in Rieger, B. (ed.) *Glottometrika 13*, Universitätsverlag Dr. N. Brockmeyer, Bochum, pp. 287-298.

Altmann, G. and Schwibbe, M. H. (1989) *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Georg Olms Verlag, Hildesheim.

Bejček, E. et al. (2013) *Prague Dependency Treebank 3.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, available at: http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3, accessed: 30 April 2022.

Berdicevskis, A. (2021) "Successes and failures of Menzerath's law at the syntactic level", in Čech, R. and X. Chen (eds.) *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, Association for Computational Linguistics, Sofia, pp. 1-16, available at: https://aclanthology.org/2021.quasy-1.2, accessed: 30 April 2022.

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

Bohn, H. (2002) "Untersuchungen zur chinesischen Sprache und Schrift", in Köhler, R. (ed.) *Korpuslinguistische Untersuchungen zur quantitativen und systemtheoretischen Linguistik*, Universität Trier, Trier, pp. 127-177, available at: https://ubt.opus.hbz-nrw.de/opus45-ubtr/frontdoor/deliver/index/docId/146/file/05_bohn.pdf, accessed: 30 April 2022.

Boroda, M. G. and Altmann, G. (1991) "Menzerath's law in musical texts", *Musikometrika*, 3, pp. 1-13.

Buk, S. (2014) "Quantitative analysis of the novel Ne spytavšy brodu by Ivan Franko", *Speech and Context: International Journal of Linguistics, Semiotics and Literary Science*, 1(6), pp. 100-112, available at: https://ibn.idsi.md/sites/default/files/imag_file/Quantitative%20analysis%20of%20the%20novel.pdf, accessed: 30 April 2022.

Buk, S. and Rovenchak, A. (2008) "Menzerath–Altmann Law for Syntactic Structures in Ukrainian", *Glottotheory*, 1(1), pp. 10-17, doi: 10.1515/glot-2008-0002.

Čech, R. and Mačutek, J. (2021) "The Menzerath-Altmann Law in Czech Poems by K. J. Erben", in Plecháč, P. et al. (eds.) *Tackling the Toolkit: Plotting Poetry through Computational Literary Studies*, Institute of Czech Literature of the Czech Academy of Sciences, Prague, pp. 5-14, doi: 10.51305/ICL.CZ.9788076580336.01.

Čech, R. et al. (2020) "Proč (někdy) nemíchat texty aneb Text jako možná výchozí jednotka lingvistické analýzy (Why not to mix texts (sometimes): The text as a possible default unit of linguistic analysis)", *Naše řeč (Our Language)*, 103(1-2), pp. 24-36.

Che, W., Li, Z. and Liu, T. (2010) “LTP: A Chinese Language Technology Platform”, in Liu, Y. and Liu, T. (eds.) *Coling 2010: Demonstrations*, Coling 2010 Organizing Committee, Beijing, pp. 13-16, available at: https://aclanthology.org/C10-3004/, accessed: 30 April 2022.

Chen, H. (2018) “Testing the Menzerath-Altmann Law in the Sentence Level of Written Chinese”, *Open Access Library Journal*, 5(e4747), pp. 1-5, doi: 10.4236/oalib.1104747.

Chen, H. and Liu, H. (2016) “How to Measure Word Length in Spoken and Written Chinese”, *Journal of Quantitative Linguistics*, 23(1), pp. 5-29, doi: 10.1080/09296174.2015.1071147.

Chen, H. and Liu, H. (2019) “A quantitative probe into the hierarchical structure of written Chinese”, in Chen, X. and Ferrer-i-Cancho, R. (eds.) *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, Association for Computational Linguistics, Paris, pp. 1-8, doi: 10.18653/v1/W19-7904.

Chen, H. and Liu, H. (2022) “Approaching language levels and registers in written Chinese with the Menzerath–Altmann Law”, *Digital Scholarship in the Humanities*, pp. 1-15, doi: 10.1093/llc/fqab110.

Chen, H., Liang, J. and Liu, H. (2015) “How Does Word Length Evolve in Written Chinese?”, *PLOS ONE*, 10(9), pp. 1-12, doi: 10.1371/journal.pone.0138567.

Cramer, I. M. (2005a) “Das Menzerathsche Gesetz”, in Köhler, R., Altmann, G. and Piotrowski, R. G. (eds.) *Quantitative Linguistik / Quantitative Linguistics: Ein internationales Handbuch / An International Handbook,* Walter de Gruyter, Berlin, New York, pp. 659–688.

Cramer, I. (2005b) “The Parameters of the Altmann-Menzerath Law”, *Journal of Quantitative Linguistics*, 12(1), pp. 41-52, doi: 10.1080/09296170500055301.

de Marneffe, M.-C. et al. (2021) “Universal Dependencies”, *Computational Linguistics*, 47(2), pp. 255-308, doi: 10.1162/coli_a_00402.

Grotjahn, R. and Altmann, G. (1993) “Modelling the Distribution of Word Length: Some Methodological Problems”, in Köhler, R. and Rieger, B. B. (eds.) *Contributions to Quantitative Linguistics: Proceedings of the First International Conference on Quantitative Linguistics, QUALICO, Trier, 1991*, Springer, Dordrecht, pp. 141-153, doi: 10.1007/978-94-011-1769-2_9.

Grzybek, P. and Stadlober, E. (2007) “Do we have problems with Arens’ law? A new look at the sentence-word relation”, in Grzybek, P. and Köhler, R. (eds.) *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, De Gruyter Mouton, Berlin, Boston, pp. 205-218, doi: 10.1515/9783110894219.205.

Hammerl, R. and Sambor, J. (1993) *O statystycznych prawach jezykowych*, Zakład Semiotyki Logicznej Uniwersytetu Warszawskiego, Warszawa.

Hinkle, D. E., Wiersma, W. and Jurs, S. G. (2003) *Applied Statistics for the Behavioral Sciences*, 5th edn, Houghton Mifflin, Boston.

Hou, R. et al. (2017) “A Study on Correlation between Chinese Sentence and Constituting Clauses Based on the Menzerath-Altmann Law”, *Journal of Quantitative Linguistics*, pp. 1-17, doi: 10.1080/09296174.2017.1314411.

Hou, R. et al. (2019a) “Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis”, *Glottometrics*, 45, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g45zeit.pdf, accessed: 30 April 2022.

Hou, R. et al. (2019b) “Linguistic characteristics of Chinese register based on the Menzerath—Altmann law and text clustering”, *Digital Scholarship in the Humanities*, pp. 1-13, doi: 10.1093/llc/fqz005.

Hřebíček, L. (2002b) *Vyprávění o lingvistických experimentech s textem*, Academia, Praha.

Institute of Computing Technology of Chinese Academy of Science (n.d.) *Chinese Lexical Analysis System ICTCLAS*, available at: https://github.com/NLPIR-team/NLPIR, accessed: 30 April 2022.

James, L. S. et al. (2021) “Phylogeny and mechanisms of shared hierarchical patterns in birdsong”, *Current Biology*, 31(13), pp. 2796-2808, doi: 10.1016/j.cub.2021.04.015.

Jiang, X. and Jiang, Y. (2022) “Menzerath-Altmann Law in Consecutive and Simultaneous Interpreting: Insights into Varied Cognitive Processes and Load”, *Journal of Quantitative Linguistics*, pp. 1-19, doi: 10.1080/09296174.2022.2027657.

Jiang, Y. and Ma, R. (2020) “Does Menzerath–Altmann Law Hold True for Translational Language: Evidence from Translated English Literary Texts”, *Journal of Quantitative Linguistics*, pp. 1-25, doi: 10.1080/09296174.2020.1766335.

Jin, H. and Liu, H. (2017) “How will text size influence the length of its linguistic constituents?”, *Poznan Studies in Contemporary Linguistics*, 53(2), pp. 197–225, doi: 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, pp. 7-27.

Kelih, E. (2010) “Parameter interpretation of Menzerath law: evidence from Serbian”, in Grzybek, P., Kelih, E. and Mačutek, J. (eds.) *Text and Language: Structures, Functions, Interrelations, Quantitative Perspectives*, Praesens, Wien, pp. 71-78.

Köhler, R. (1982) “Das Menzerathsche Gesetz auf Satzebene”, in Lehfeldt, W. and Strauss, U. (eds.) *Glottometrika 4*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 103-113.

Köhler, R. (1984) “Zur Interpretation des Menzerathschen Gesetzes”, in Boy, J. and Köhler, R. (eds.) *Glottometrika 6*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 177-183.

Köhler, R. (2012) *Quantitative syntax analysis*, Walter de Gruyter, Berlin, Boston.

Köhler, R. and Naumann, S. (2009) “A contribution to quantitative studies on the sentence level”, in Köhler, R. (ed.) *Issues in Quantitative Linguistics*, RAM-Verlag, Lüdenscheid, pp. 34-45.

Kułacka, A. (2008) “Badania nad prawem Menzeratha–Altmanna”, *LingVaria*, 2(6), pp. 167-174.

Laboratory for Chinese Character Research and Application (n.d.) *Chinese Character Holographic Resource Application System*, available at: https://qxk.bnu.edu.cn/#/, accessed: 30 April 2022.

Lee, J., Leung, H. and Li, K. (2017) “Towards Universal Dependencies for Learner Chinese”, in de Marneffe, M.-C., Nivre, J. and Schuster, S. (eds.) *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, Association for Computational Linguistics, Gothenburg, pp. 67-71, available at: https://aclanthology.org/W17-0408/, accessed: 30 April 2022.

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

Mačutek, J. and Rovenchak, A. A. (2011) "Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length", in Kelih, E., Levickij, V. and Matskulyak, Y. (eds.) *Issues in Quantitative Linguistics 2*, RAM-Verlag, Lüdenscheid, pp. 136-147.

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

Mačutek, J., Čech, R. and Courtin, M. (2021) "The Menzerath-Altmann law in syntactic structures revisited: Combining linearity of language with dependency syntax", in Čech, R. and Chen, X. (eds.) *Proceedings of the Second Workshop on Quantitative Syntax (Quasy, SyntaxFest 2021)*, Association for Computational Linguistics, Sofia, pp. 1-9, available at: https://aclanthology.org/2021.quasy-1.6/, accessed: 30 April 2022.

Mačutek, J., Čech, R. and Milička, J. (2017) "Menzerath-Altmann law in syntactic dependency structure", in Montemagni, S. and Nivre, J. (eds.) *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, Linköping University Electronic Press, Pisa, pp. 100-107, available at: https://aclanthology.org/W17-6513.pdf, accessed: 30 April 2022.

Mačutek, J., Chromý, J. and Koščová, M. (2018) "Menzerath-Altmann Law and Prothetic /v/ in Spoken Czech", *Journal of Quantitative Linguistics*, pp. 1-15, doi: 10.1080/09296174.2018.1424493.

Matoušková, L. (2016) "An Application of the Menzerath-Altmann Law to a Text Written in Traditional Chinese Characters", in Benešová, M. (ed.) *Text segmentation for Menzerath-Altmann law testing*, Palacký University Olomouc, Olomouc, pp. 44-70.

Matoušková, L. and Motalová, T. (2015) "An Application of the Menzerath-Altmann law to Chinese translations of the poem The Raven", *Czech and Slovak Linguistic Review*, 2, pp. 49-64.

McEnery, M., Xiao, Z. and 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), pp. 361-378, doi: 10.1093/llc/18.4.361.

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

Miller, G. A. (1956) "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information", *The Psychological Review*, 63(2), pp. 81-97.

Motalová, T. and Matoušková, L. (2014) *An Application of the Menzerath-Altmann Law to Contemporary Written Chinese*, Palacký University Olomouc, Olomouc.

Motalová, T. et al. (2013) "An Application of the Menzerath-Altmann Law to Contemporary Written Chinese", *Czech and Slovak Linguistic Review*, 1, pp. 22-53.

Nivre, J. et al. (2020) "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection", in Calzolari, N. et al. (eds.) *Proceedings of the 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, pp. 4034–4043, available at: https://aclanthology.org/2020.lrec-1.497, accessed: 30 April 2022.

Osborne, T. (2019) *A Dependency Grammar of English: An introduction and beyond*, John Benjamins, Amsterdam, Philadelphia.

Pelegrinová, K., Mačutek, J. and 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), pp. 405-414, doi: 10.2478/jazcas-2021-0037.

Pinyin4j (n.d.) Pinyin4j, available at: http://pinyin4j.sourceforge.net/, accessed: 30 April 2022.

Python-pinyin (2022), Mozillazg, available at: https://github.com/mozillazg/python-pinyin, accessed: 30 April 2022.

Roukk, M. (2007) “The Menzerath-Altmann law in translated texts as compared to the original texts”, in Grzybek, P. and Köhler, R. (eds.) *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, De Gruyter Mouton, Berlin, Boston, pp. 605-610, doi: 10.1515/9783110894219.605.

Schwibbe, M. H. (1984) “Text- und wortstatistische Untersuchungen zur Validität der Menzerathschen Regel”, in Boy, J. and Köhler, R. (eds.) *Glottometrika 6*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 152-176.

Semple, S., Ferrer-i-Cancho, R. and Gustison, M. L. (2021) “Linguistic laws in biology”, *Trends in Ecology & Evolution*, pp. 1-14, doi: 10.1016/j.tree.2021.08.012.

Shahzad, K., Mittenthal, J. E. and Caetano-Anollés, G. (2015) “The organisation of domains in proteins obeys Menzerath-Altmann’s law of language”, *BMC Systems Biology*, 9(44), doi: 10.1186/s12918-015-0192-9.

Sherrod, P. H. (2005) *NLREG Version 6.3 (Advanced)*.

Stalph, J. (1989) *Grundlagen einer Grammatik der sinojapanischen Schrift*, Ph.D. Dissertation.

Stave, M. et al. (2021) “Optimisation of morpheme length: a cross-linguistic assessment of Zipf’s and Menzerath’s laws”, *Linguistics Vanguard*, 7(s3), doi: 10.1515/lingvan-2019-0076.

Sun, F. and Caetano-Anollés, G. (2021) “Menzerath–Altmann’s Law of Syntax in RNA Accretion History”, *Life*, 11(6), pp. 1-18, doi: 10.3390/life11060489.

Sun, P. and Shao, Y. (2021) “Verification of Menzerath-Altmann law in Different Chinese Registers based on corpus”, in *2021 2nd International Conference on Artificial Intelligence and Education (ICAIE)*, IEEE, Dali, pp. 34-39, doi: 10.1109/ICAIE53562.2021.00014.

Švarný, O. and Uher, D. (2001) *Hovorová čínština: Úvod do studia hovorové čínštiny – 2. díl (Spoken Chinese: Introduction to Study of Spoken Chinese – Part 2)*. Palacký University Olomouc, Olomouc.

Taylor, J. R. (ed.) (2015) *The Oxford Handbook of the Word*, Oxford University Press, Oxford.

Teupenhayn, R. and Altmann, G. (1984) “Clause length and Menzerath's law”, in Boy, J. and Köhler, R. (eds.) *Glottometrika 6*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 127-138.

Torre, I. G., Dębowski, Ł. and Hernández-Fernández, A. (2021) “Can Menzerath’s law be a criterion of complexity in communication?”, *PLoS ONE*, 16(8), doi: 10.1371/journal.pone.0256133.

UD Chinese GSDSimp (2021), Universal Dependencies, available at: https://universaldependencies.org/treebanks/zh_gsdsimp/index.html, accessed: 30 April 2022).

Valente, D. et al. (2021) “Linguistic laws of brevity: conformity in Indri indri”, *Animal Cognition*, 24, pp. 897-906, doi: 10.1007/s10071-021-01495-3.

Wang, L. and Čech, R. (2016) "The impact of code-switching on the Menzerath-Altmann Law", *Glottometrics*, 35, pp. 22-27, available at: https://www.ram-verlag.eu/wp-content/uploads/2018/08/g35zeit.pdf, accessed: 30 April 2022.

Wee, L.-H. and Li, M. (2015) "Modern Chinese Phonology", in Wang, W. S.-Y. and Sun, C. (eds.) *The Oxford Handbook of Chinese Linguistics*, Oxford University Press, New York, pp. 474-489.

Wimmer, G. et al. (2003) *Úvod do analýzy textov (Introduction to Text Analysis)*, Slovenská akademie věd (Slovak Academy of Sciences), Bratislava.

Wong, T.-sum et al. (2017) "Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank", in Montemagni, S. and Nivre, J. (eds.) *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, Linköping University Electronic Press, Pisa, pp. 266–275, available at: https://aclanthology.org/W17-6530/, accessed: 30 April 2022.

Zeman, D. et al. (2017) "CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies", in Hajič, J. and Zeman, D. (eds.) *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, Vancouver, pp. 1-19, doi: 10.18653/v1/K17-3001.

Zeman, D. et al. (2021) *Universal Dependencies 2.9*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, available at: http://hdl.handle.net/11234/1-4611, accessed: 30 April 2022.

汉字信息词典 (Dictionary of Chinese Character Information) BCC 语料库 - 北京语言大学 (BCC Corpus – Beijing Language and Culture University), 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: 2 December 2021.

文林 Wénlín Software for Learning Chinese: Version 4.0.2 (2011), *Wenlin Institute, Inc*, available at: http://www.wenlin.com.

List of publications

Book and book chapters

Motalová, T. and Spáčilová, L. (2014) *An Application of the Menzerath-Altmann Law to Contemporary Written Chinese*, Palacký University Olomouc, Olomouc.

Motalová, T. and Matoušková (Spáčilová), L. (2016) “An Application of the Menzerath-Altmann Law to Contemporary Written Chinese”, in Benešová M. (ed.) *Menzerath-Altmann Law Applied*, Palacký University Olomouc, Olomouc, pp. 87–120.

Motalová, T. and Schusterová, D. (2016) “Menzerath-Altmann Law – Analyses of Short Stories Written by Chinese Author”, in Benešová, M. (ed.) *Menzerath-Altmann Law Applied*, Palacký University Olomouc, Olomouc, pp. 71–116.

Motalová, T. (2022) “The Menzerath-Altmann Law in Syntactic Relations of Chinese Language Based on the Universal Dependencies (UD)”, in Yamazaki, M., Sanada, H., Köhler, R., Embleton, S., Vulanović, R. and Wheeler, E. (eds.) *Quantitative Approaches to Universality and Individuality in Language*, De Gruyter Mouton, (in print).

Articles

Matoušková, L. and Motalová, T. (2015) “An Application of the Menzerath-Altmann law to Chinese translations of the poem The Raven”, *Czech and Slovak Linguistic Review*, 2, pp. 44-61.

Kovaľová, J., Lavička, M., Matoušková, L., Motalová, T-, Vicherová Schusterová, D., Szokalová, K. (2016) “Výzkum převodu zahraničních proprií do čínštiny (The adaptation of foreign proper names in Chinese)”, *Nový Orient (New Orient)* 71(3), pp. 34-44.

Schusterová, D., Motalová, T. and Wang F. (2017) “Menzerath-Altmann定律: 理论贡献与局限 [Menzerath-Altmann Law: Contributions and Limitations of the Theory]”, in Hu, Z. (ed.) *语言学研究. 第二十一辑 [Linguistic Research. Volume 21]*. Peking: Higher Education Press, pp. 45-58.

Review

Motalová, T. (2015) Review of Čínská obchodní konverzace by Slaměníková, T. and Guo, Y., *Nový Orient (New Orient),* (70)3, pp. 74-76.