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 | 2022

Declaration

I declare that this dissertation is the result of my individual and independent research. All of the sources consulted have been properly cited.

Olomouc, 8 August 2022

Signature

Acknowledgement

First and foremost, I thank my supervisor Radek Čech for guidance and endless support – without his supervision, this work would not have been possible.

I thank Ján Mačutek, Jiří Milička and Xinying Chen for the time I stole from them to answer my questions and, of course, for their answers.

I thank Martina Benešová for her supervision during my first years of PhD. studies.

I thank my colleagues at work for their patience and kindness.

I thank Katka and Yulia for sharing the burden.

I thank Michal, bearing my 'writing mode' was not always easy.

I thank my family and all the people around who did not give up on me, even when I disappeared for a while and mostly said 'No. I will not join'.

Contents

Introduction

Language is viewed by synergic linguistics (Köhler, 2005) as a self-organizing and self-regulating system which interacts with its environment while adapting itself to it and manifesting itself through various phenomena we can observe. Based on our knowledge, or more precisely, theoretical frameworks, universal hypotheses can be derived, tested and combined into a network of laws that eventually can explain these language phenomena (Köhler, 2012). One of these universal hypotheses is the Menzerath-Altmann law.

The 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. For example, the shortest words are expected to consist of the longest syllables having the most sounds. The average number of sounds per syllable decreases while the number of syllables in words increases until the longest words consist of the shortest syllables having the least sounds. 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). The tendency for the negative correlation between lengths of two language units was observed by Menzerath (1954) and later mathematically formalized by Altmann (1980). Nowadays, it is known as the Menzerath-Altmann law and 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, and even beyond the borders of linguistics (e.g. proteins, animal communication). However, particular language units (e.g. word) are drawing more attention from researchers than others (e.g. phrase). Hence, knowledge about their behaviour in relation to the law is imbalanced. 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). Although a unit in one position might behave in accord with the law, its behaviour might change if its position is switched over to the other. 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). Another issue arises with regard to the evaluation of results. The law is corroborated if the agreement between empirically obtained results and theoretical results predicted by the law, or more precisely, by its model, reaches a certain degree. However, researchers do not agree on a minimum threshold at which the law becomes corroborated and follow different rules of thumb. Generally speaking, the research on the law often shows a lack of consensus on applied methods, which hinders appropriate comparison of achieved results and blurs the overall picture for the scope of the law's validity (e.g. Köhler, 2012; Berdicevskis, 2021).

We aim to address these challenges within this thesis. Therefore, we set several general and language-specific objectives. 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. Except for peripheral units, 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. 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, or in other words, their determinations 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 two studies focusing on Chinese (Chen and Liu, 2019, 2022) already applied the law to a hierarchy of language units, both left the phrase level out of the analysis. 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, especially when its hierarchical position changes. Finally, both the studies (also in Chen and Liu, 2016) 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 (e.g. frequency) or the specific word length distribution in Chinese can be regarded as the boundary conditions for the law.

In Chapter One, we introduce the Menzerath-Altmann law in detail – we shortly describe its discovery and then shift focus towards its interpretations that have been made so far and the challenges its application faces. In Chapter Two, we provide an overview of studies and their results achieved by testing given linguistic levels in various languages, including Chinese. In Chapter Three, we present the methodology. We describe a language material under analysis, determine each language unit and introduce individual unit combinations to which the law is applied. In Chapter Five, we present the results. The final chapter draws conclusions.

1 Menzerath-Altmann law

The Menzerath-Altmann law deals with a relation between language units which are positioned in a vertical hierarchy according to their size – with a bigger unit on a higher level A while consisting of smaller units of a lower level B and with the unit on the B level while consisting of smaller units of a lower level C (Hřebíček, 2002a, p. 25). As Hřebíček (2002b, pp. 59-60) explains, such a structure resembles Russian dolls where each element is bigger than all smaller elements and smaller than all bigger elements at the same time (even though language units are allowed to be equal in their size). The relation between these units is negatively correlated – with an increase in the length of the A unit measured as a sum of B units, the mean length of the B unit measured in the C units decreases. The calculation of the mean length can be illustrated with a simple equation $b = \frac{c}{a}$, where c is, for example, a sum of phonemes in a word, a is a sum of syllables in the word, and, finally, b is the mean size of the syllable in the phonemes in this word.

1.1 The law's discovery

Observations on relations between lengths of respective units were (probably) firstly made in phonetics, where the duration of a syllable was brought into focus (Altmann and Schwibbe, 1989, p. 60). In this connection, studies published at the end of the 19th and beginning of the 20th century are usually mentioned. For example, Sievers (1901) pointed out that the duration of syllables tends to be shorter if a speech act consists of more syllables and vice versa.[1] Grégoires (1899) observed changes in the duration of the same vowel, which shortens with longer words.[2] Other studies subsequently appeared to confirm or question such observations (for their overview, see Altmann and Schwibbe, 1989, p. 60).

Menzerath was, however, the first who formulated his observations in the form of laws regulating relations between lengths of sounds, syllables and words and tried to interpret them (Cramer, 2005a, p. 660). The earlier work – Menzerath and de Oleza (1928) – presents findings from an experiment on approximately 1500 Spanish words. Based on the results, the authors outlined general laws describing quantitative changes in lengths of tested units – firstly, the mean duration of a sound shortens with longer words measured either in the number of sounds or syllables; secondly, the mean duration of a syllable gets shorter with an increasing number of syllables in a word; lastly, mean duration of a word increases with an increasing number of sounds or syllables in words (Menzerath and de Oleza, 1928, pp. 68-76).[3] In 1954, Menzerath published another work where he corroborated a particular lawful relationship – "[d]ie relative Lautzahl nimmt mit steigender Silbenzahl ab"[4] (Menzerath, 1954, p. 100) – for more than 20k

[1] E.g. in Altmann and Schwibbe (1989, p. 60), Best (2007, p. 92).

[2] E.g. in Cramer (2005b, pp. 41-42), Kułacka (2009a, p. 55).

[3] E.g. in Cramer (2005a, p. 660), Best (2007, pp. 88-93), Best and Rottmann (2017, p. 100).

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

German words. Moreover, he generalized the findings as follows "je größer das Ganze, um so kleiner die Teile!"[5] (Menzerath, 1954, p. 101) and interpreted it as a result of economy rules. Despite Menzerath's appeal, mathematical formalisation and further research were not carried out until almost three decades later in an article published by Altmann (1980).

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.

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

Altmann corroborated the law's validity for Indonesian morphemes and English words – both being the constructs to syllables measured in phonemes – by using the formula (2). The third experiment applied the formula (1) and showed that the lengths of Bachka-German words and syllables (measured in a unit of time) are also in accordance with the law (Altmann, 1980, pp. 6-8).[6]

Thanks to the contributions of both the authors – Menzerath and Altmann – the law is acknowledged and well-known as the Menzerath-Altmann law[7].

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

[6] Altmann (1980, pp. 7-8) analysed a spoken German dialect of Bachka, a geographical area located in the Pannonian basin.

[7] Coined by Hřebíček (1990b).

1.2 The law's interpretation

The Menzerath-Altmann law, among other quantitative linguistic laws, is considered to be one of the universal hypotheses of synergic linguistics (e.g. Köhler, 1999, 2005, 2012). Synergic linguistics assumes the language to be a self-organising and self-regulating system optimally adapting itself to its environment (Köhler, 1993, p. 41). Based on the modelling, synergic linguistics "can be used to set up universal hypotheses by deduction from theoretical considerations, to test them, to combine them into a network of laws and law-like statements, and to explain the phenomena observed" (Köhler, 2012, p. 169). As Vulanović and Köhler (2005, p. 283) explain, such a hypothesis (or a law) derived from a model for a language mechanism and revealing its details is representational, or in other words, a grey or white box. On the other hand, a law which only describes a relationship between two quantities – or phenomena – without revealing details about its internal mechanism is phenomenological, or in other words, a black box. Several attempts have been made to shed light on the mechanism behind the Menzerath-Altmann law.

Menzerath interpreted his conclusion "je größer das Ganze, um so kleiner die Teile"[8] (Menzerath, 1954, p. 101) as a result of economy rules which ensure manageability of the whole (1954, p. 101). Similarly, Altmann (1980, p. 5) associated the law with the principle of least effort or another unknown principle that balances lengthening and shortening tendencies.

Schwibbe (1984, 1989) explored the linkage between the law and noise generated over the course of transmission of information through a channel. The longer the information, the higher the amount of noise in the channel and the higher the degree of activation of the central nervous system (CNS). In order to compensate for this burden and ensure the reliability of the transmitted information, the processing system shortens the information by splitting it into smaller segments. Schwibbe (1989) tested his assumption on normal letters and suicide notes. The latter showed a greater shortening tendency which probably balances a higher amount of noise and a higher degree of activation of CNS caused by extreme stress conditions.

Köhler (1984; also in Vulanović and Köhler, 2005; Köhler, 2012) assumed that language is sequentially processed in a so-called register which might be associated (or even identified) with short-term memory. The register functions as storage on each level, firstly, for a currently processed constituent and, secondly, for a result of analysis (or synthesis) – i.e. structural information – which, according to Köhler, carries information about connections among constituents of a language construct. The limited capacity of the register regulates storage distribution – the more structural information the construct needs for its constituents, the less storage is available for the constituents themselves. As a consequence, the construct length has its upper limit. The combination of the plain information (=constituents of a given construct) and the structural information resulting in the construct being, in fact, larger than the total of its constituents has been further developed by Milička (2014, p. 89).

Kułacka (2009a) made a direct link between the law and working (also called immediate or short-term) memory by putting it into the context of the capacity theory of comprehension (proposed by Just and Carpenter, 1992). According to the theory (Just and Carpenter, 1992, p.

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

123), each element to be comprehended has its so-called activation level. If the activation level is above a certain value, it becomes a part of the working memory. However, working memory has its upper threshold – if a required amount of activation for comprehension is higher than the threshold, working memory is re-organised and old elements displaced. As Kułacka (2009a) explained, when processing language units with their activation levels, if a greater amount of activation is taken by the construct, less space can be used by its constituents. Or in other words, the higher the complexity of the construct, the lower the complexity of its constituents.

Apart from Köhler (1984) and Kułacka (2009a), there are other studies which connected the law to limits of short-term memory, or more precisely, to the 'magical number plus or minus two' proposed by Miller (1956). These studies evaluated whether a constituent length under analysis (or its determination) is in accord with this number representing an amount of information which we are able to process in short-term memory. For example, Jiang and Ma (2020) evaluated a clause measured in words or Mačutek, Čech and Courtin (2021) the clause measured in linear dependency segments. Jin and Liu (2017, p. 217) used Miller's number to point out that an informal and conversational nature of a sample of fiction prose obeys the limited span of short-term memory and, as a result, clause lengths have a lower number of words on average. As Jiang and Ma (2020, p. 19) added, the short-term memory limits might be boundaries for a reasonable information flow in a language. Similarly, Araujo, Benevides and Pereira (2020, p. 43) argued that the concept of a more complex construct having simpler constituents is in accord with Miller's limit of short-term memory.

Generally speaking, limits of the cognitive capacity of a human mind and its overload are often seen as a cause of the law (e.g. Jin and Liu, 2017; Jiang and Ma, 2020; Jiang and Jiang, 2022). It is believed that greater exploitation of the capacity leads to faster release of this cognitive burden resulting in a greater shortening tendency of the constituent lengths. For example, Jiang and Ma (2020, pp. 17-18) revealed that texts translated into a target language show a greater decrease in clausal lengths (being constituents to a sentence) than texts written directly in such a language. The authors connected the extra load to double-processing of a language material – decoding texts from a source language and encoding them into a target language. The analysis by Jiang and Jiang (2022, pp. 7-12) showed that transcriptions of simultaneous interpreting, i.e. interpreting which transforms a message into a target language while the message is being produced (Strazny, 2005, p. 535), have a faster decreasing tendency of clausal lengths (being constituents to a sentence) compared to transcriptions of consecutive interpreting, i.e. interpreting which converts the message into the target language after the message is produced (Strazny, 2005, pp. 534-535). The authors argued that simultaneous interpreting exploits the cognitive capacity to a larger extent than consecutive interpreting.

Hou et al. (2017) looked at the scope of the law's validity from a perspective of the difference between writing and speaking – their results showed that the law was mainly corroborated for written formal texts contrary to texts of conversational nature. When producing a text, the former requires planning, whereas the latter lacks the need since conversations are spontaneous and cannot be changed afterwards. Hence, the authors made a link between the validity of the law and samples of the written language style while excluding samples of the spoken language style from the law's scope.

Fenk-Oczlon and Fenk (1995) came up with time limits that might constrain the lengths of language units. The authors primarily focused on analysing a relationship between syllable and sentence lengths even though the syllable is not considered a direct constituent to the sentence in the menzerathian view (for more detail, see Chapter 1.4). The relationship was tested on the same collection of several declarative sentences translated into almost 30 languages. The authors firstly calculated the overall mean lengths of the sentences for each language and revealed that the lengths are mainly in the range of 7 ± 2 syllables, i.e. Miller's number (1956). Secondly, when calculating the mean sizes of syllables in phonemes, results showed that languages with a lower mean length of the sentences tend to have a higher mean size of syllables (i.e. higher complexity) and vice versa. The authors believe that the regulation of the syllable complexity is a consequence of the properties of the language system, which ensures a constant and economical flow of linguistic information by using the Menzerath-Altmann law. Keeping the length of the sentences in a certain range while adequately regulating the syllable sizes enables to meet a limited time window for perception or production of the sentence.

The law has also been discussed in connection with breathing and lung capacity. The need to inhale might force a sound producer or a speaker to shorten constituents in their lengths (Torre, Dębowski and Hernández-Fernández, 2021, p. 2). Physical units determined by the breathing rhythm of humans have already been tested by Rothe-Neves, Marques Bernardo and Espesser (2017), who applied the law to a speech segment uttered in one stream, and by Torre et al. (2019) and Hernández-Fernández et al. (2019), who opted for a breath group determined by breaks for inhalation. Following the interpretation of the authors, the results corroborated the law, and the corroboration led Hernández-Fernández et al. (2019, p. 12) to conjecture that the law's fundamentals might originate from acoustics.[9]

Hřebíček introduced a completely different view on the law. In his works (e.g. 1994, 1995, 1997, 2002b), the author explored a link between the law and a fractal. The fractal is understood as a structure whose parts resemble the whole. Or in other words, the structure is self-similar (Hřebíček, 1994, p. 84). Hřebíček (1994, p. 86) believed that a fractal character of a language stems from the self-similarity of language constructs and its constituents whose relationships are in accord with the law. "[T]he movement up or down the ladder of the language levels results in the sort of symmetry which represents similarity; the mutually similar items are located inside each other. This is the characteristic property of fractals" (Hřebíček, 2002a, p. 20). Afterwards, the potential connection between the law and the fractal was further developed by Andres (e.g. 2010; 2014), who mathematically formalised the self-similarity dimension for the language fractal, called a degree of semanticity, based on isomorphism between formulas of the law and the self-similarity dimension of the mathematical fractal (Andres, 2017).

[9] If reviewing the results, the achieved goodness-of-fit between a model and data expressed by the coefficient of determination R^2 might not be regarded as satisfactory – in the case of English $R^2 = 0.7$ (Torre et al., 2019, p. 17), in the case of Spanish $R^2 = 0.84$ and in the case of Catalan $R^2 = 0.47$ (Hernández-Fernández et al., 2019, p. 10). For comparison, Mačutek and Wimmer (2013, p. 233) mention a value of 0.90 and higher to indicate a satisfactory fit.

1.3 The law's controversy

Even though we might have some clues about the mechanism behind the law, the law still faces difficulties with the interpretation of parameters integrated into its models. The degree of interpretability depends on a particular parameter in question. Nevertheless, it is generally understood that the lack of a solid linguistic interpretation makes parameters just numbers generated by models fitted to particular language data (e.g. Meyer, 2002, p. 69) and makes the models mathematically descriptive rather than explanatory (Mačutek and Wimmer, 2013, p. 236).

The parameter a is usually described as a value on the y-axis where a fitting curve starts if the model (3) is applied.[10] 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. $y = y_1 x^b$ (e.g. Köhler, 1984, p. 180; Teupenhayn and Altmann, 1984, pp. 128-129; Cramer, 2005b, p. 50; Kelih, 2010, p. 75). Andres et al. (2012b, p. 6) used $a = \frac{y_1}{e^c}$ instead of the parameter a in the formula $y = ax^{-b}e^{cx}$, leading to its modified version, i.e. $y = y_1 x^{-b} e^{c(x-1)}$. However, the replacement complicates the parameter's interpretability in this case. Köhler (1984, pp. 180-181) assumed that the value of the parameter a is specific to a particular language and text but later specified (2012, p. 147) that its dependency on an analysed linguistic level overrides the influence of language, text or author. This dependency was shown by Cramer (2005b, pp. 46-50), who re-analysed data obtained by other researchers on various linguistic levels. Čech et al. (2020, p. 33) showed that the parameter a reaches similar values on the word level for five texts of different types and authors (i.e. $2.55 \leq a \leq 2.64$). Nevertheless, the influence of genre, text and author has been under discussion too. Teupenhayn and Altmann (1984, pp. 128-129) or Altmann and Schwibbe (1989, p. 43) drew such a connection to stylistics while analysing the sentence level. Čech and Mačutek (2021, p. 11) confirmed that values of the parameter a significantly differ on the word level for poetic and prosaic texts (based on a statistical test). Kułacka (2010, pp. 261-266) arrived at a similar conclusion when she analysed empirical values y_1, i.e. mean sizes of clauses (in words) of mono-clausal sentences.[11] Her results confirmed the influence of a text type and a language – scientific texts and English texts showed a greater value of y_1 than literary texts and Polish texts.[12] Kułacka's conclusion of the empirical

[10] Hřebíček (1995, p. 56) and later Andres (2010, p. 110) also mentioned a connection between the parameter a and the number of hapax legomena, however, without any further details.

[11] Based on the results obtained from preliminary analysis, Kułacka (2010, p. 261) determined $y_1 = 10$ to be a threshold value for tested text types, i.e. $y_1 > 10$ for scientific texts and $y_1 < 10$ for literary texts.

[12] Kułacka (2010, p. 262) explained a higher value of y_1 in English by its rather analytical nature leading to the usage of more words compared to Polish which uses affixes to express the same meaning due to its inflectional nature.

value being below a certain threshold for literary texts (2010, p. 262) was also corroborated on the same linguistic level by Jiang and Ma for corpora of short stories (2020, p. 13).[13]

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, 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). In Köhler's view (1984, pp. 178-181), the b parameter also reflects a degree of the increase in structural information, which adequately changes with an increasing construct length, while Milička (2014, p. 89) suggested that it represents a mean length of structural information. Köhler (1982, p. 110) firstly assumed that the parameter b is a language and possibly text specific but later argued (2012, p. 147) that its value mainly depends on a linguistic level under analysis which was again corroborated by Cramer (2005b, p. 50). Similarly, the parameters b obtained from samples of different Slavic languages (Kelih, 2008, pp. 19-20) or monolingual text types (Kelih, 2010, p. 74) did not significantly differ (based on a statistical test), implying that "a common statistical mechanism seems to organise the relation of word and syllable length" (Kelih, 2010, p. 74). Čech and Mačutek (2021, pp. 11-12) came to the similar conclusion that the syllable lengths decrease with the same 'speed' on the word level since differences between poetic texts of one author and prosaic texts of another were not significantly different with respect to the parameters b (based on statistical tests). On the other hand, Čech et al. (2020, p. 33) showed that the parameter b is influenced on the word level by a text type – two presidential speeches had values of the parameter b close to each other while other texts, each of a different text type, differed. However, no statistical test was carried out because of the limits of the tested sample. The influence of the text type was also demonstrated by Kułacka (2010, pp. 266-267) on the sentence level – lower values resulting in steeper slopes of fitting curves emerged in scientific texts while tested languages (English and Polish) did not considerably influence the value. However, Kułacka's assumption of the parameter b being above a certain threshold for literary texts (2010, p. 266)[14] was not corroborated for all samples of short stories analysed by Jiang and Ma (2020, p. 13). The authors concluded that "b might be more sensitive than a if used to capture typological differences" (Jiang and Ma, 2020, p. 16). Teupenhayn and Altmann (1984, p. 129) suggested that a value of the parameter b which is outside a confidence interval (i.e. a range of values that a parameter has with a certain degree of probability, e.g. Dekking et al., 2005) might indicate a text being produced under abnormal circumstances (with regard to psychology or psycholinguistics). As for the relation to the language fractal, Hřebíček (1997, p. 39) interpreted the parameter b as the inverse similarity dimension.[15] Andres and Benešová (2011, 2012) calculated the self-similarity dimension of the language fractal – called degree of semanticity – as a reciprocal mean of the parameters b, which were obtained from linguistic

[13] Values of the parameter a were in accord with Kułacka's threshold (2010, p. 261) for literary texts, i.e. $a < 10$.

[14] Kułacka (2010, p. 266) determined a threshold value for the parameter b to be equal to -0.1, i.e $b < -0.1$ for scientific texts and $b > -0.1$ for literary texts.

[15] As Hřebíček explains (1997, p. 39), the law represents an inverse formulation of the similarity between a whole and its part, i.e. it expresses the similarity between the mean part and the whole.

levels tested on a sample (considered to be the language fractal if all the levels corroborate the law).

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). Cramer (2005b, p. 51) corroborated the systematic connection between the parameters by correlation and variance analyses and assumed as well as Teupenhayn and Altmann (1984, p. 129) that the value of the parameter b could be estimated from a value of the parameter a. The dependency of the parameter b on the parameter a was also supported by Altmann and Schwibbe (1989, p. 43 and pp. 57-58), who 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). Based on their findings, it appears that the parameter b depends on the parameter a, and their values correlate with each other. Nevertheless, the predictability of the parameter b remains an open question. Köhler (1984, pp. 180-181; Köhler, 1989, p. 111; Vulanović and Köhler, 2005, p. 283) even assumed that the parameters should be in the linear relation, ideally, if the constituent and the structural information fully exploit the register. Following the assumption, Kelih (2010, p. 76) modelled empirical values y_1 and the parameters b by a linear equation (i.e. $b = -0.2869 \times SyL_1 + 0.6528$, where SyL_1 is a mean syllable length of monosyllabic words) and confirmed such a tendency on a word level for Serbian texts of different types and their mixture. Similar results were brought by Hou et al. (2019a, p. 37) on the clausal level. However, two questions arise. First, how we can interpret the parameters used in the linear formula (Mačutek and Rovenchak, 2011, p. 141). Second, under which condition does such a linear relation emerge because it has not been confirmed, for example, by Mačutek and Rovenchak (2011, p. 141) for Ukrainian and Indonesian canonical word form types[16]. Recently, the values of both the parameters have been used for cluster analyses which revealed a tendency of samples to cluster together according to the text types to which they belong. Xu and He (2018, pp. 10-11) showed that corpora of spoken academic discourse clustered together as well as corpora of written academic discourse. Hou et al. (2019b, p. 8) confirmed a cluster for corpora of conversations while a corpus of news stayed separated. Two clusters representing two types of interpreting (i.e. simultaneous and consecutive) can be found in Jiang and Jiang (2022, pp. 12-14). Chen and Liu (2022, p. 8) also revealed a similar clustering tendency of two text types (press and scientific texts). All the authors (Xu and He, 2018, p. 10; Hou et al., 2019b, p. 11; Chen and Liu, 2022, p. 8; Jiang and Jiang, 2022, p. 13) supported the idea of using the parameters for differentiation of text types. Mačutek, Čech and Milička (2017, p. 105) even addressed that the parameters combined with the dependency syntax might be exploited for authorship or language typology analyses.

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).

[16] The canonical word form consists only of two types of phonemes – vowels and consonants (Mačutek and Rovenchak, 2011, p. 136).

Moreover, it appears, based on its value, that the exponential part e^{-cx} is more relevant to lower linguistic levels (e.g. phonetic or word level) while being irrelevant to higher ones (e.g. syntactic level) (Vulanović and Köhler, 2005, p. 283; Andres et al., 2012b, p. 6; Köhler, 2012, p. 148). Andres (2014, p. 31) even raised an objection to the exponential part, which is somewhat artificial and lacks a solid linguistic ground in his view.

The controversial interpretability of the parameters closely relates to the absence of consensus on the choice of a particular model with regard to tested data (mentioned already by Cramer, 2005a, p. 633). In general, more parameters usually lead to a better fit. However, if an additional parameter lacks a plausible interpretation, a model with a smaller number of parameters should be preferred (e.g. Grzybek, 1999, p. 74; Köhler, 2012; p. 53; Milička, 2014, p. 96). As Köhler pointed out, it is "a trade-off between the two criteria – improvement of the goodness-of-fit on the one hand and number of parameters on the other" (2012, p. 53).

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).

This increasing tendency was already expected by Altmann (1980)[17] and later called a second (Torre et al., 2019, p. 14; Torre, Dębowski and Hernández-Fernández, 2021, p. 2) or reverse (Tanaka-Ishii, 2021, p. 11) regime of the law. It usually occurs in the form of two phenomena across results yielded by studies. Firstly, it is expected that the longest constituent appears together with the shortest construct and a fitting curve starts decreasing from its head. However, some studies showed that the constituent reaches its highest value with the second shortest construct and the peak of the fitting curve is consequently shifted (see Figure 1), e.g. for physical units[18] in Torre et al. (2019), on the syntactic level in Hou et al. (2017), Hou et al. (2019b), Berdicevskis (2021), Tanaka-Ishii (2021), on the word level in Altmann and Schwibbe (1989), Lehfeldt and Altmann (2002), Benešová, Faltýnek and Zámečník (2015), Mačutek, Chromý and Koščová (2018), Čech et al. (2020). The phenomenon even led, for example, Kraviarova and Zimmermann (2010) and Torre, Dębowski and Hernández-Fernández (2021) to exclude one-constituent constructs from analyses.

[17] We remind the reader that the possible occurrence of such a tendency led Altmann to reformulate his first verbal expression of a monotonical decrease to "[t]he length of the components is a function of the length of language constructs" (Altmann, 1980, p. 3).

[18] Torre et al. (2019, p. 2 and p. 16) applied the law to breath groups determined by pauses in speech for breathing and words being measured in three different units (characters, phonemes or time units). The shifted peak of the fitting curve appeared when the word was measured in characters and phonemes.

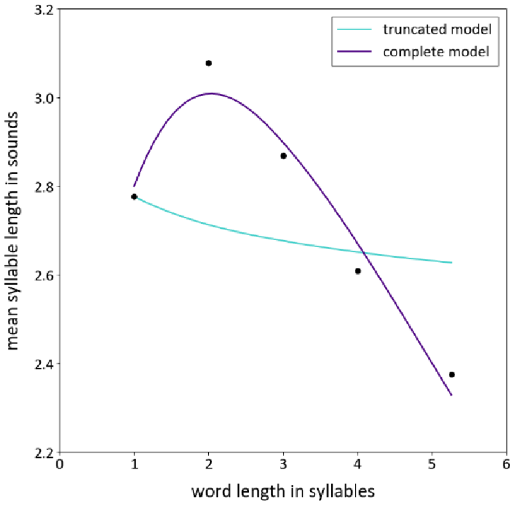


Figure 1. The example of the law's second regime in the form of the peak of a fitting curve being shifted.

Next, the second or reverse regime mainly occurs with the longest constructs. The constituent lengths first decrease as expected and then start oscillating in an upward trend while the construct lengths continue increasing. Hence, a fitting curve rises in its tail (see Figure 2). Such an unusual behaviour diverging from the menzerathian tendency has occurred, for example, on the syntactic level in Heups (1983), Hug (2004), Jin and Liu (2017), Hou et al. (2019b), Berdicevskis (2021), Tanaka-Ishii (2021), Chen and Liu (2022) and on the word level in Torre, Dębowski and Hernández-Fernández (2021). Hug (2004, p. 9) even posed the question of whether the scope of the law is not limited rather to the shortest constructs.

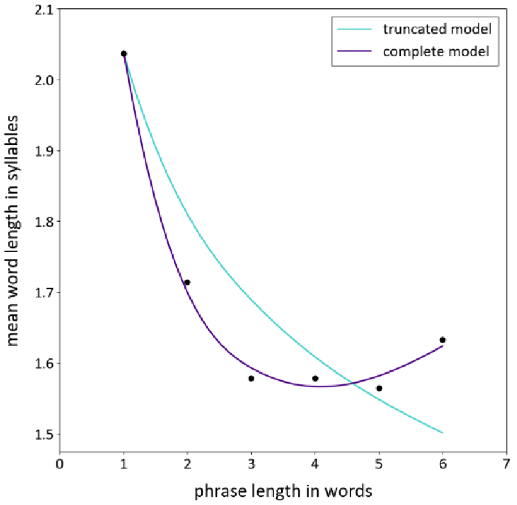


Figure 2. The example of the law's second regime in the form of the increase in a curve's tail.

It is noteworthy that the considerable fluctuation of the constituent lengths is mainly linked to higher variability of a sample resulting in low frequencies of the longest constructs to which the constituents belong (e.g. Altmann, 1980, p. 5, Altmann and Schwibbe, 1989, p. 37; Mačutek, Čech and Milička, 2017, p. 104). For this reason, researchers usually either omit such observations or apply the method of the weighted average (i.e. the construct and constituent lengths are pooled together and weighted according to their frequency). The frequency minimum the construct must reach otherwise is treated with one of the methods mentioned above varies across studies. The observations were omitted if their frequency f was $f < 5$ (e.g. on the syntactic level in Jin and Liu, 2017; Xu and He, 2018; Jiang and Jiang, 2022; on the word level in Wimmer et al., 2003; Mačutek and Rovenchak, 2011), $f \leq 10$ (e.g. on the syntactic level in Köhler, 1982; Heups, 1983; Bohn, 2002; Benešová and Čech, 2015; Mačutek, Čech and Milička, 2017; on the word level in Bohn, 2002; Mačutek, Chromý and Koščová, 2018; Rujević et al., 2021) or even higher (e.g. on the syntactic level $f < 50$ in Berdicevskis, 2021; on the word level $f < 20$ in Milička, 2014; $f < 25$ in Torre, Dębowski and Hernández-Fernández, 2021). There are also a number of studies which did not follow any rule of thumb as the previous works but omitted only particular construct lengths with a low frequency (e.g. on the syntactic level in Teupenhayn and Altmann, 1984; Kułacka, 2009b; on the word level in Altmann and Schwibbe, 1989; Grzybek, 2000; Köhler, 2002; Lehfeldt and Altmann, 2002; Buk and Rovenchak, 2007; Kelih, 2010; Kraviarova and Zimmermann, 2010). The weighted average was applied by Mačutek, Čech and Courtin (2021) on the syntactic level for $f < 10$ and by Čech and Mačutek (2021) on the word level for $f < 5$. It is noteworthy that researchers also raised the question of whether other factors contribute to these fluctuations (e.g. Mačutek, Čech and Milička, 2017, p. 104). Kelih (2010, p. 73) associated the irregular behaviour with long lengths of words in general, while Mačutek and Rovenchak (2011, p. 139) pinpointed (but did not test) compound words not being possibly driven by the menzerathian mechanism. Regarding the sentence level, the text size and its degree of regulation might be taken into account as another factor. Jin and Liu (2017) achieved an excellent fit between a model and data when they applied the law to a collection of Chinese microblogs, i.e. posts whose size is restricted to 140 Chinese characters per each.[19] The authors did not omit any sentence length being in the range from one to seven clauses (all reaching the frequency $f \geq 5$), and none of the data points considerably fluctuated from a fitting curve. The authors believe that the results reflect the self-organisational and self-regulatory properties of the language, which responds to the size restriction and correspondingly adapts the lengths of sentences and, consequently, the lengths of clauses measured in words (Jin and Liu, 2017, pp. 216-217). This property might not be so noticeable if the text size is not restricted.

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).

[19] The fit was expressed by the coefficient of determination R^2, i.e. $R^2 = 0.998$ (Jin and Liu, 2017, p. 215).

To fit data with alternative models for the law is no exception across studies (e.g. Lehfeldt and Atmann, 2002; Buk and Rovenchak, 2007; Kułacka and Mačutek, 2007; Mačutek and Rovenchak, 2011; Milička, 2014; Altmann and Gerlach, 2016; Best and Rottmann, 2017; Rujević, 2021). However, we will not go into detail since the work does not aim to be a complex theoretical analysis of the law and its mathematical formalisation but rather to be an analysis of its application to particular language data fitted by Altmann's models, i.e. the complete model $y = ax^b e^{-cx}$ and the truncated model $y = ax^b$ (with the parameter a being substituted by empirically obtained lengths).

Finally, to illustrate the point of how the choice of the model influences results, two examples can be used. First, when Benešová, Faltýnek and Zámečník (2015) fitted their data with the standard model (3), i.e. $y = ax^b$, the goodness-of-fit was low.[20] Since the constituent lengths showed the second (or reverse) regime, Mačutek, Chromý and Koščová (2018) re-fitted their data with the general model (2), i.e. $y = ax^b e^{-cx}$, and yielded a considerable improvement of the original fit.[21] The study by Rujević et al. (2021) can serve as a second example – the general model (2), $y = ax^b e^{-cx}$, fitted to word tokens of four languages did not show good results, whereas an alternative model derived by the authors achieved an excellent fit. However, at the cost of the higher number of parameters, i.e. $y(x) = ax^{b+c\log x} e^{-dx}$, where $y(x)$ is a constituent length of a given construct x and a, b, c and d are the parameters.

1.4 The law's (in)validity

The law has been corroborated by a number of studies which applied the law to various language materials and language units (see Chapter 2). 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 (Wilde and Schwibbe, 1989; Ferrer-i-Cancho and Forns, 2010; Li, 2012; Ferrer-i-Cancho et al., 2014; Nikolaou, 2014; Sun and Caetano-Anollés, 2021), or animal communication of birds (Favaro et al., 2020; James et al., 2021), or primates (Gustison et al., 2016; Fedurek, Zuberbühler and Semple, 2017, Gustison and Bergman, 2017; Heesen et al., 2019; Clink, Ahmad and Klinck, 2020; Clink and Lau, 2020; Huang et al., 2020; Watson et al., 2020; Valente et al., 2021).[22]

However, there are also results which rejected the law, e.g. for the syntactic level in Bohn (1998, 2002), Roukk (2007), Buk and Rovenchak (2008), Kułacka (2009b), Sanada (2016), Hou et al. (2017), for the word level in Buk (2014), Chen and Liu (2016, 2019, 2022), Mačutek, Chromý and Koščová (2018), Čech and Mačutek (2021), for particular primate duets in Clink and Lau (2020), gorillas' close-call sequences (Huang et al., 2020). As Köhler (2012, p. 175) pointed

[20] Based on the coefficient of determination R^2, i.e. $R^2 = 0.5710$ and $R^2 = 0.6253$ (Benešová, Faltýnek and Zámečník, 2015, p. 45).

[21] $R^2 = 0.8940$ and $R^2 = 0.9618$ accordingly (Mačutek, Chromý and Koščová, 2018, p. 4).

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

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). To illustrate the point, an example of the boundary condition for the Menzerath-Altmann law might be monosyllabic words in old Russian before the elimination of specific vowels (Altmann and Lehfeldt, 2002, p. 36). In the menzerathian view, the shortest construct is expected to be composed of the longest constituents on average. However, the syllable structure in old Russian allowed the length of the monosyllabic words to be only up to two phonemes which imposed limitations on the law to come into force (Altmann and Lehfeldt, 2002, p. 36). This might relate to a conjecture that the law 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; Kułacka, 2009b, p. 27). Similarly, Sanada (2016, pp. 267-269) argued that the construct, i.e. clause, might be restricted to have only a certain number of its constituents, i.e. arguments of a predicate. Such a restriction might cause a low variability of the construct lengths and consequently mean constituent lengths being rather constant and independent of the construct. A limit imposed by a text size was suggested by Čech and Mačutek (2021, p. 8) based on results 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, for example, by Heups (1983, p. 119), Teupenhayn and Altmann (1984, pp. 129-130), Altmann and Schwibbe (1989, p. 38), Hug (2004, p. 9) and Cramer (2005a, p. 663). 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), e.g. a poet who chooses particular – shorter – syllables due to euphony (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 (as addressed by Grotjahn and Altmann, 1993, for modelling of the word length distribution)[23]. In the following paragraphs, we do not aim to provide an exhaustive probe into these issues but rather to outline the complexity which arises when the Menzerath-Altmann law is applied.

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. Hence, samples should achieve homogeneity to the greatest possible extent. "[T]exts will often be more homogeneous if they are shorter, less revised and written more spontaneously" (Best and Rottmann, 2017, p. 39). Additionally, a sample which is homogenous,

[23] The authors also discussed the problem of modelling and explanation. For more details, see Grotjahn and Altmann (1993).

for example, for testing the frequency of phonemes might not be homogenous enough for testing sentence lengths and vice versa. When sampling, a property of a unit in question should also be taken into account since the homogeneity of the same sample does not have to be applied to more properties, or in other words, it is not transferrable between them (Altmann, 1992, p. 291). Altmann (1992, pp. 290-291) and Grotjahn and Altmann (1993, pp. 143-144) suggest first analysing closed text parts (e.g. individual chapters) to test "whether the parameters of the model are stationary" (Grotjahn and Altmann, 1993, pp. 143-144) and then to analyse the whole text if the parts are homogenous. Grotjahn and Altmann (1993, pp. 143-144) further explain this approach by different factors influencing text production and, consequently, units and their properties throughout the text (e.g. word length). Similarly, Wimmer et al. (2003, p. 89) argue that a long text might be produced with interruptions causing changes, hence, it might be divided into sections (e.g. chapters), otherwise, the whole text should be analysed. The analysis of the whole text is preferred, for example, by Best and Rottmann (2017, p. 40), who regard the text as an individual stylistic unit or by synergic linguistics, which considers the text to be an organised and balanced system produced under certain initial and (hypothetically) stable conditions (Uhlířová, 1995, p. 10). Similarly, Hřebíček (2002b, p. 43) emphasises a context which forms language units of various linguistic levels into a text, or more precisely, a coherent structure with a clear beginning and end while not even being interrupted by non-textual elements (e.g. pictures).

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). Altmann (1988, pp. 155-156) assumes that selections from one text would follow a model rather than selections from several texts unless they are homogenous. From the perspective of synergetic linguistics, either systematic or random text selections might distort text features or even cause their loss (Uhlířová, 1995, p. 10). Best and Rottmann (2017, p. 40) consider a mixture of texts to be a mixture of different styles violating homogeneity. On the one hand, a mixed sample can cause a mechanism not to reveal itself and, consequently, a tested hypothesis to be rejected, on the other hand, it can also cause the mechanism to be amplified more than in individual texts (Čech, 2020, pp. 26-28).

Several examples can illustrate the double-edged nature of text mixing. The disagreement between the Menzerath-Altmann law and data on the sentence level is associated with the heterogeneity of literary text types. They are primarily written works, but due to their frequent inclusion of dialogues, they also approximate the spoken form of a language. A conversational property of such samples tends to shorten clauses on average which might a) prevent the Menzerath-Altmann law from coming into play and b) lead to worse results (Kułacka, 2009b, p. 27; Jin and Liu, 2017, p. 217; Hou et al., 2017, pp. 10-11). Different speakers representing different speech styles might be another factor which amplifies the degree of heterogeneity and brings about unsatisfactory results (Altmann and Schwibbe, 1989, p. 61; Mačutek, Chromý and Koščová, 2018, p. 4, when reviewing results by Benešová, Faltýnek and Zámečník, 2015, who applied the law to a dialogue of four different speakers). On the contrary, Kelih (2010, p. 74; 2012, p. 210) showed that the heterogeneity of a monolingual sample containing several text types does not considerably lower the goodness-of-fit for word types. Čech et al. (2020) and Jiang and Ma (2020) even achieved the best fitting results for a mixture of all texts under analysis. The former study tested the law on texts of different types and showed that, on the one hand, the individual texts yielded a worse (but still satisfactory) fit, on the other hand, they differed in values of the parameters (b, c) and courses of fitting curves which might be specific to the text type or the author (Čech et al. 2020, pp. 32-35).[24] In the case of the latter study, the goodness-of-fit between the model and the data depended on the sampling. As mentioned above, the mixed sample achieved the best fit.[25] When zooming into this sample containing four collections of Lu Xun's short stories translated by four different translators, we find out that the law was corroborated only for two collections.[26] If we zoom again into one of these collections, we see that only three of eight short stories translated by the same translator corroborated the law (Jiang and Ma, 2020, p. 15, 24).

Studies on the Altmann-Arens' law can also demonstrate the impact of heterogeneity (Grzybek and Stadlober, 2007; Grzybek, Stadlober and Kelih, 2007; Grzybek, Kelih and Stadlober, 2008). The Altmann-Arens' law deals with a positive correlation between the lengths of the construct and its indirect constituents. Altmann (1983) interpreted Arens' observation (1965) of the simultaneous increase in the word and sentence lengths as a reverse tendency of the Menzerath-Altmann law if a linguistic level is skipped. However, Grzybek and Stadlober (2007, p. 208) re-analysed Arens' data and revealed poor fitting results unless the data were pooled.[27] As the authors addressed, the Menzerath-Altmann law is of intra-textual nature, i.e. being related to the internal structure of a text (or group of texts), while Arens calculated the mean length of words and sentences in each text and analysed "the relationship between these means across different texts" (Grzybek and Stadlober, 2007, p. 209). This led the authors to question whether the Altmann-Arens' law is a consequence of the Menzerath-Altmann law and, therefore, of the intra-textual too, or it is the inter-textual law being applicable across text types (Grzybek and Stadlober, 2007, pp. 208-209; Grzybek, Stadlober and Kelih, 2007, pp. 3-4). Firstly, Grzybek and Stadlober (2007) tested the Altmann-Arens' law on the inter-textual level following the Arens' approach. The Arens' data being mixed with another dataset of two text types yielded even worse results. The authors preliminarily concluded that the Altmann-Arens' law might be related only to particular text types that sufficiently vary (since the pooled Arens' data showed a good fit). Continuing to analyse the inter-textual level, Grzybek, Stadlober and Kelih (2007, pp. 5-6) showed only a weak relationship between means of the word and sentence lengths when testing each chapter of a Russian novel or 199 Russian texts of six text types (being analysed as the whole and as individual text types). In the authors' view, the samples lacked sufficient length variability. The later study (Grzybek, Kelih and Stadlober, 2008, p. 119) shifted the focus towards

[24] Values of both the parameters were close to each other in the case of two texts representing presidential speeches (Čech et al., 2020, p. 33). However, differences between the parameters were not statistically tested because of the limits of the sample.

[25] We review the results based on the coefficient of determination R^2 following the standard of the thesis, i.e. $R^2 \geq 0.90$.

[26] The third collection was slightly below the standard, i.e. $R^2 = 0.8841$.

[27] Altmann (1983) used F-test, while Grzybek and Stadlober (2007) tested the data using the coefficient of determination R^2. Due to the high variability of insufficiently large data, Grzybek and Stadlober (2007, pp. 212-213) eventually pooled the means a) into classes including five observations and b) based on intervals of sentence lengths to make the tendency more apparent.

the intra-textual level and revealed that as far as a sample is large and heterogeneous enough, the expected reverse tendency appears.[28] The authors corroborated the tendency for a sample of several text types (drama, comment, letters, literary texts), its partial version without literary texts and literary texts themselves. The authors concluded that the menzerathian tendency in the form of Altmann-Arens' law seems to work for the external textual heterogeneity (mixture of text types) and internal textual heterogeneity (literary texts being heterogenous enough due to the inclusion of various textual elements, e.g. dialogues and comments (Grzybek, Kelih and Stadlober, 2008, p. 119).

The interrelation of linguistic properties relates to the frequency which influences the manifestation of the Menzerath-Altmann law. This issue primarily concerns lower linguistic levels (e.g. the word) because of the higher probability that the same unit can occur more than once within a sample. The higher the linguistic level (e.g. clause or sentence), the lower the probability. Such a frequency reflects a unit usage, i.e. deals with 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).[29] 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 fully manifest itself (in a similar manner discussed in Hug, 2004; Mikros and Milička, 2014; Pelegrinová, Mačutek and Čech, 2021; Rujević et al., 2021; Stave et al., 2021). The biasing impact of the Brevity law can be diminished by analysing the unit types whose constituents tend to have higher mean lengths than constituents of the unit tokens. We can take monosyllabic words in Ukrainian (Buk, 2014, pp. 107-108) and German (Best and Rotmann, 2017, pp. 103-104) as examples. Their syllable lengths equal 3.32 or 3.37 phonemes in the types and 2.30 or 2.88 phonemes in the tokens accordingly. Menzerath (1954) was aware of this influence, and due to his interest in the structure of languages, he examined the types to avoid the prevalence of words with high frequency. Similarly, Altmann and Schwibbe (1989, p. 51) argued in favour of counting a unit only once, i.e. its types, as well as Kelih, who explained the choice by the nature of the law being a "construction mechanism" (2008, p. 14). As Stave et al. concluded, "Menzerath's Law is expected to be due to an intrinsic trade-off between the components and the carrier, and not to the frequency of usage of the specific carrier" (2020, p. 4). On the other hand, Chen and Liu (2022, p. 5) suggested that the analysis of the tokens might contribute to the recognition of text types.

[28] In the case of a collection of sentences which fulfil particular conditions. The authors excluded the shortest sentences whose words showed a monotonic decrease in their lengths, the longest sentences and sentences with a frequency equal to or lower than 30, which showed a higher variance in word lengths (Grzybek, Kelih and Stadlober, 2008, pp. 115-119).

[29] 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).

Let us review the results obtained when the Menzerath-Altmann law was applied to word tokens and types being the construct to syllables (or characters in Chinese). In the case of the word tokens, the corroboration of the law was yielded by Wimmer et al. (2003), Milička (2014) and Rujević et al. (2021) – the goodness-of-fit achieved a satisfactory value.[30] However, Wimmer et al. (2003) fitted a model only to three word lengths, while Milička (2014) and Rujević et al. (2021) used an alternative formula. Some studies corroborated the law but not for all samples under analysis (Mačutek, Chromý and Koščová, 2018[31]; Galieva, 2021[32]; Torre, Dębowski and Hernández-Fernández, 2021[33]) or not for all data points (Kraviarova and Zimmermann, 2010[34]). Lastly, there are studies in which analysis of the word tokens did not bring corroborating results at all (Alekseev, 1998; Motalová and Matoušková, 2014; Benešová, Faltýnek and Zámečník, 2015; Chen and Liu, 2016, 2019, 2022, when testing the word measured in Chinese characters as mentioned above).[35] When it comes to the word types, the situation is more straightforward. Almost all studies yielded corroboration of the law based either on the apparent menzerathian decreasing tendency (Altmann and Schwibbe, 1989; Buk and Rovenchak, 2007, although when using an alternative formula; Dinu and Dinu, 2009; Altmann and Gerlach, 2016; Araujo, Benevides and Pereira, 2020) or a satisfactory goodness-of-fit (Menzerath, 1954[36]; Bohn, 2002; Grzybek, 1999, 2000; Köhler, 2002; Kelih, 2008, 2010, 2012; Mačutek and Rovenchak, 2011, when also fitting an alternative formula to data). Only Čech and Mačutek (2021) did not corroborate the law for all samples.[37] Finally, four studies simultaneously tested both – tokens and types – while showing only the types corroborating the law (Alekseev, 1998; Buk, 2014; Mikros and Milička, 2014; Best and Rottmann, 2017).[38] However, Buk (2014) and Best and Rottmann (2017)[39] fitted the data with an alternative formula, while Mikros and Milička (2014) just tested the monotonic decrease of the constituent lengths, which was violated by disyllabic words in the case of the tokens.

The Menzerath-Altmann law operates with the concept of the construct and constituent standing for units of measurement. As Altmann (1983; also Köhler, 1982, p. 109; Altmann and Schwibbe, 1989, pp. 46-48; Cramer, 2005a, pp. 633-634, Köhler and Naumann, 2009, p. 38) pointed out, the negative correlation between lengths of the construct and the constituent only

[30] Expressed by the coefficient of determination R^2 in accordance with $R^2 \geq 0.90$.

[31] For eight out of 10 texts with respect to $R^2 \geq 0.90$.

[32] For three out of six if following the same standard of $R^2 \geq 0.90$.

[33] Showing the negative correlation only for half of tested samples representing 21 languages.

[34] The decreasing tendency was confirmed if the authors excluded one-constituent constructs from the analysis.

[35] The goodness-of-fit did not follow the standard of $R^2 \geq 0.90$, or the decreasing tendency was considerably violated.

[36] Menzerath's data (1954) was later re-analysed by Fenk, Fenk-Oczlon and Fenk (2005), who yielded the fit in accord with $R^2 \geq 0.90$.

[37] Two out of 13 samples did not reach $R^2 \geq 0.90$. Nevertheless, the goodness-of-fit of one poem was slightly below the standard, i.e. $R^2 = 0.883$ (Čech and Mačutek, 2021, p. 9), while the second might be too short for the law to come into force.

[38] In case of the goodness-of-fit being in accord with $R^2 \geq 0.90$ (if R^2 applied) or the presence of the apparent decreasing menzerathian trend.

[39] The tokens yielded a fit slightly below the standard, i.e. $R^2 = 0.88$ (Best and Rottmann, 2017, p. 103).

emerges if the immediately adjacent units are tested, or in other words, the levels are not skipped. Despite the different approach, Altmann (1983) followed up Arens' findings (1965) of the sentence and the word lengths being positively correlated and associated this increasing trend with Menzerath's law, or more precisely, with its general form – "[t]he length of the components is a function of the length of language constructs" (Altmann, 1980, p. 3).[40] For example, if a sentence length increases along with the decrease in the length of a clause, then the clause length decreases along with the increase in the length of its direct constituents, i.e. words. Hence, leaving the clause out should result in the reverse tendency – the sentence length increases along with the increase in the word length, or in other words, the word length is a function of the sentence length. Nevertheless, testing this reverse relationship faces an issue in obtaining sufficient data points, especially on higher linguistic levels, because the construct lengths measured in indirect constituents (e.g. sentence in words) can vary to a larger extent than being measured in its direct constituents and the trend might not appear (Köhler and Naumann, 2009, pp. 38-39; Köhler, 2012, p. 108).[41]

Apart from the studies by Grzybek and Stadlober (2007), Grzybek, Stadlober and Kelih (2007), Grzybek, Kelih and Stadlober (2008), Grzybek (2010) and Grzybek (2013) which showed the positively correlated relationship between the sentence and the word lengths either for pooled or sufficiently large and heterogeneous data, we can find more results obtained when various linguistic levels were skipped – either on the construct or the constituent level. Motalová and Matoušková (2014), Benešová and Birjukov (2015), Birjukov (2016), and Motalová and Schusterová (2016) measured the clause (a segment between selected punctuation marks being called a parcellate or an intercomma) indirectly in Chinese or Japanese characters (roughly corresponding to a syllable and being measured in components). The clause in the position not only of the construct but also of the constituent to the sentence led to similar results. Data points were more or less scattered around the fitting curve without any predominant tendency (being slightly increasing, decreasing or even constant). The following two triplets are other examples of skipping linguistic levels when measuring the constituent – the clause as the constituent of the sentence being indirectly measured in syllables and the phrase as the constituent of the clause being indirectly measured in morphemes. The former triplet yielded the inverse menzerathian tendency (Buk and Rovenchak, 2008), while in the latter case, the tendency was rather constant (Sanada, 2016). The skipping also appeared on a word level. The word length indirectly measured in grapheme or phoneme was chosen as the constituent to the phrase (Berdicevskis, 2021) or the clause (Hug, 2004; Berdicevskis, 2021[42]) and revealed both the positive and negative correlation. The studies mentioned above have in common that they

[40] Arens (1965) analysed only two coordinates per text as described above, while the menzerathian approach considers all categories of the construct length in a text.

[41] We remind the reader that Grzybek, Kelih and Stadlober (2008) included only sentence lengths with a frequency > 30 in their analysis to reduce data variance.

[42] On the one hand, Berdicevskis (2021, p. 11) addressed that morphemes or syllables might be preferred as the measurement units of the word. On the other hand, the author argued that word length in phonemes or graphemes might be highly correlated with the word length in morphemes or syllables (as he illustrated his point with the positive correlation between graphemic and morphemic lengths of words in Swedish).

skipped only one of the measurement units (direct or indirect constituent to the construct). However, there are also studies which skipped both. Such an example can be the triplet composed of the sentence, word and grapheme (phoneme or phone), which was tested by Hřebíček (2002a), Hug (2004) and Berdicevskis (2021). The results again vary – Hřebíček (2002a) showed fluctuation in the constituent lengths while Hug (2004) and Berdicevskis (2021) detected both correlations. Although skipping a linguistic level mainly leads to at least ambiguous or even worse results, Chen and Liu (2022) demonstrated that leaving a unit out can bring a better goodness-of-fit between a model and data. Initially, the authors did not corroborate the law for word tokens measured directly in Chinese characters and indirectly in components. Hence, the authors decided to leave the Chinese character out and apply the law to the triplet of word, component and stroke, which led to a considerably increased fit.[43] Nevertheless, going one level above, the clause being combined with the word measured in the components yielded almost the same results as the word measured in Chinese characters.[44]

The Menzerath-Altmann law seems sensitive to a choice of measurement units and their mutual distance. If the construct and its constituent are not close (or far) enough, the analysis brings various results. To illustrate the point, we can use the results brought by Berdicevskis (2021). When starting with the sentence level, the negative correlation between sentence, word and grapheme was confirmed for 26 out of 78 languages. The number of languages increased to 68 when including the clause (sentence, clause and word) but dropped to 38 languages when changing the word to the phrase (sentence, clause, phrase). When going one linguistic level lower, lengths of the clause as the construct and lengths of the word as its constituents measured in graphemes were negatively correlated only in 12 languages. However, when the clause was measured in phrases and words, the number of languages raised to 58 even though including the phrase on the higher linguistic level yielded worse results.

Another question arises about how skipping linguistic levels relates to engaging time as the unit of measurement. For example, the combination of the word measured in phonemes and the phoneme measured in seconds corroborated the law despite the fact that a linguistic level (e.g. syllable or morpheme) was skipped (Hernández-Fernández et al., 2019; Torre et al., 2019).[45]

The last issue to be discussed here is the evaluation of results. Earlier studies (e.g. Altmann, 1980; Köhler, 1982; Heups, 1983; Schwibbe, 1984; Teypenhayn and Altmann, 1984; Altmann and Schwibbe, 1989) used F-test which statistically test sampled data against a null hypothesis predicting a zero correlation between variables (Grotjahn, 1992, p. 129). However, its use for language data was later criticised because the significance of the F-test leading to acceptance of a model might be caused by sample size (e.g. Grotjahn, 1992, pp. 124-125). The goodness-of-it between the model and data started to be commonly evaluated by the coefficient of determination R^2 which reflects the degree of agreement between empirical and

[43] From 0.1625 to 0.8982 based on values of the coefficient of determination R^2 (Chen and Liu, 2022, p. 5).

[44] $R^2 = 0.7657$ and $R^2 = 0.7477$ accordingly (Chen and Liu, 2022, p. 5).

[45] At least for English (Torre et al., 2019, p. 17) and Spanish (Hernández-Fernández, 2019, p. 11) based on the coefficient of determination R^2, i.e. $R^2 = 0.90$. The fit for Catalan was considerably lower, i.e. $R^2 = 0.75$ (Hernández-Fernández, 2019, p. 11).

theoretical values (Kelih, 2008, p. 17; used e.g. by Prün, 1994; Hřebíček, 1995; Grzybek, 1999; Bohn, 2002; Wimmer et al., 2003; Roukk, 2007; Kelih, 2008; Mačutek and Rovenchak, 2011; Köhler, 2012; Milička, 2014; Benešová and Čech, 2015; Sanada, 2016; Mačutek, Čech and Milička, 2017; Xu and He, 2018; Chen, 2018; Jiang and Ma, 2020; Mačutek, Čech and Courtin, 2021; Jiang and Jiang, 2022). 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. According to Andres et al. (2012a, p. 15), the adequate goodness-of-fit is achieved when $R^2 \geq 0.70$. The minimum value for good results starts at $R^2 \geq 0.80$ for Best and Rottmann (2017, p. 101) or at $R^2 \geq 0.85$ for Grzybek and Stadlober (2007, p. 208) and Kelih (2008, p. 17). Mačutek and Wimmer et al. (2013, p. 233) even refer to the rule of thumb in the form of $R^2 \geq 0.90$. Some authors use a scale for the interpretation of their results. Acceptable results start with $R^2 \geq 0.70$ (Jin and Liu, 2017; Xu and He, 2018; Hou et al., 2019a, 2019b; Jiang and Ma, 2020; Jiang and Jiang, 2022) or $R^2 \geq 0.75$ (Chen, 2018; Chen and Liu, 2019, 2022), the good results at $R^2 \geq 0.80$ (Chen, 2018; Chen and Liu, 2019, 2022) or at $R^2 \geq 0.85$ (Jin and Liu, 2017; Jiang and Jiang, 2022) and excellent results at $R^2 \geq 0.90$ (Jin and Liu, 2017; Xu and He, 2018; Chen, 2018; Chen and Liu, 2019, 2022; Hou et al., 2019a, 2019b; Jiang and Jiang, 2022). The lack of consensus on the threshold blurs an overall picture regarding the scope of the law's validity. Spearman's rank correlation coefficient is another method applied to test the menzerathian relationship between the construct and its constituent, and used, for example, by Kułacka (2009b), Berdicevskis (2021), Torre, Dębowski and Hernández-Fernández (2021). Its value ranges between -1 and 1 for the negative and positive correlation respectively and its significance is tested by the *p*-value (Torre, Dębowski and Hernández-Fernández, 2021, p. 8). However, the coefficient tests only the correlation and not the fit of the model to the data. As Berdicevskis pointed out, it is "not really informative for the languages with clear non-monotonic patterns" (2021, p. 8), which occurrence is no exception as discussed above in connection with the second or reverse regime of the law. Some studies opted for other methods, but their usage is more or less peripheral. Hence, we will not go into further details.

2 Menzerath-Altmann law on language units

The law has been widely applied to various linguistic levels and their combinations. However, the chapter primarily summarises those findings related to linguistic levels tested by the thesis (see Table 1). As for alternative triplets which have been analysed beyond the scope of this work, we provide their brief overview only if the same construct (i.e. sentence, clause, syntactic phrase, word or character) is analysed and measured in its possible neighbouring units (studies which tested triplets where linguistic levels were apparently skipped were already introduced in Chapter 1.4). When summarising the studies, we always mention language material under analysis due to its influence on results (as addressed in Chapter 1.4) and follow interpretations provided by authors. Where studies achieved the corroboration of the law using an alternative formula proposed by authors, we add this information to the overview. Otherwise, the models derived by Altmann (1980) were applied. If the coefficient of determination R^2 is used for the evaluation of the goodness-of-fit between models and data, due to the lack of a consensus on its minimal value needed for the law's corroboration (as discussed in Chapter 1.4), 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). The chapter is divided into subchapters according to the construct in question. We start with the sentence even though there are works which go above this level and treat the sentence as the constituent, e.g. Hřebíček (1990a, 1995, 1997) used the sentence as a measurement unit for so-called hrebs or semantic aggregates, Grzybek (2013) as the constituent to a chapter, Motalová and Matoušková (2014) and Benešová and Birjukov (2015) as the constituent to a paragraph. Each subchapter introduces results achieved for various languages excluding Chinese, towards which the focus is shifted afterwards.

Table 1. Overview of the linguistic levels under analysis by the thesis.

Construct	**Direct constituent**	**Indirect constituent**
Sentence	Clause	Word
	Syntactic phrase	Word
	Clause	Syntactic phrase
Clause	Word	Character/syllable
	Syntactic phrase	Word
Syntactic phrase	Word	Character/syllable
Word	Character	Component
	Character	Stroke
	Syllable	Sound
Character	Component	Stroke

2.1 The sentence as the construct

2.1.1 The sentence across languages

The determination of the sentence is not always sufficiently addressed within studies (Kułacka, 2009b; Xu and He, 2018; Jiang and Jiang, 2022) or not addressed at all (Köhler, 1982; Schwibbe, 1984, 1989; Teupenhayn and Altmann, 1984; Köhler and Naumann, 2009; Köhler, 2012). If it is, utilised approaches vary. On the one hand, there is a common ground in determining sentence borders based on punctuation marks. On the other hand, studies do not agree on the details of the determination. Benešová and Čech (2015) used only a full stop. Other studies also considered a question mark and an exclamation mark, and the determination was usually further specified. Authors extended the selection of the punctuation marks, e.g. by a colon and a semicolon (Hug, 2004) or by an ellipsis[46] (Roukk, 2007). An additional rule can also condition the determination, e.g. Heups (1983) determined the sentence based on the capitalisation of its first grapheme (even after the colon). Some authors combined both, e.g. studies of Buk and Rovenchak (2008) and Jiang and Ma (2020) implemented the rule of capitalisation while the former added the ellipsis as the sentence-final mark and the latter the ellipsis, a dash and brackets. Even though Sanada (2016) used punctuation, she did not provide details of selected marks. Last but not least, some authors relied on the sentence determination provided by an annotation scheme of language material (Mačutek, Čech and Milička, 2017; Tanaka-Ishii, 2021; Berdicevskis, 2021; Mačutek, Čech and Courtin, 2021).

2.1.1.1 The clause as the constituent

The choice of the clause as the constituent of the sentence and the word as the constituent of the clause prevails among studies. It gives rise to the hypothesis – the longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in words.

As for the determination of the clause, or more precisely, the number of the clauses in the sentence, authors usually count finite verbs (Köhler, 1982; Heups, 1983; Teupenhayn and Altmann, 1984; Roukk, 2007; Benešová and Čech, 2015; Xu and He, 2018; Jiang and Jiang, 2022). Nonetheless, this approach is not the only operationalisation of the clause which can be found. Buk and Rovenchak (2008) developed an algorithm which identified the clause in Ukrainian based on different verb forms (excluding infinitives), predicative words, punctuation marks and conjunctions. An algorithm designed by Köhler and Naumann (2009) for the German language automatically detected three types of clauses – finite, infinite and verbless – while using punctuation and conjunctions as indicators of potential clausal boundaries.[47] Jiang and Ma (2020)

[46] In the realm of punctuation being chosen as unit boundaries, the ellipsis strictly denotes the punctuation mark composed of three or six dots.

[47] The authors checked manual and automatic determination of the clause on five texts and achieved 90-95% success rate (Köhler and Naumann, 2009, p. 38).

analysed the finite and infinite clauses operationalised as a sequence of words with subject and predicate between selected punctuation marks (comma, semicolon and colon). Hug (2004) determined the clause – being called a group – solely by punctuation, i.e. comma and dash. Similarly, Schwibbe (1984) processed the clause as a sequence of words inserted between two punctuation marks. However, the author did not further specify their selection; moreover, he conditioned the clausal length to be greater than two words (without any justification). The approach to the clause while using an annotation scheme of language material was adopted by Berdicevskis (2021). Last but not least, there are studies which did not provide (sufficient or any) details on the clause determination (Schwibbe 1989; Kułacka, 2009b; Köhler, 2012).

The word was mainly processed as a sequence of characters between two spaces (Heups, 1983; Hug, 2004; Roukk, 2007; Buk and Rovenchak, 2008; Benešová and Čech, 2015; Jiang and Ma, 2020). Berdicevskis (2021) exploited the annotation scheme of language material with minor adjustments, while Xu and He (2018) and Jiang and Jiang (2022) used parsing or tagging tools for word determination. Other studies did not address the determination of the word (Köhler, 1982; Schwibbe, 1984, 1989; Teupenhayn and Altmann, 1984; Köhler and Naumann, 2009; Kułacka, 2009b; Köhler, 2012).

The hypothesis was corroborated by empirical data from several languages. As regards English, Köhler (1982) did not reject the hypothesis for a sample of sentences selected from different texts (based on an F-test with the apparent menzerathian decreasing tendency)[48], Teupenhayn and Altmann (1984) for each sample of sentences from individual texts (based on an F-test) and Kułacka (2009b) for each excerpt from five out of seven books of literary and scientific text types (based on Spearman's correlation coefficient). The corroborating results were also achieved when testing text collections. Xu and He (2018) applied the law to corpora of spoken and written academic discourse, their mixed sample, and to a corpus of play and television scripts ($R^2 \geq 0.90$ was reached in all the cases). Jiang and Ma (2020) tested a corpus of English short stories and a corpus of English translations of Chinese short stories (however when reviewing their results in the light of the standard of $R^2 \geq 0.90$, the law would be corroborated only for the corpus of translations and for its two collections out of four if analysed separately, R^2 of the third collection was slightly below this standard, i.e. $R^2 = 0.8841$). Finally, Jiang and Jiang (2022) analysed corpora of transcriptions representing simultaneous and consecutive interpreting from Chinese to English (the standard of $R^2 \geq 0.90$ was met).

In the case of German, the corroboration of the hypothesis was yielded by analysing samples of sentences collected from book chapters (Köhler, 1982, based on an F-test with the apparent menzerathian decreasing tendency), samples of sentences from individual texts (Teupenhayn and Altmann, 1984, based on an F-test), four separate German texts (Köhler, 2012, even though only three of them would meet the standard of $R^2 \geq 0.90$), a sample of newspaper articles (Köhler and Naumann, 2009, who presented the result only in the form of a graph depicting the apparent menzerathian decreasing tendency, which, in the authors' view, corresponded to results based on manual processing of the clause published by other studies)

[48] The re-analysis by Köhler (2012) showed the coefficient of determination R^2 in accord with $R^2 \geq 0.90$.

and a corpus containing different text types, i.e. news, letters, novels, legal and scientific texts, (Heups, 1983, based on an F-test with the apparent menzerathian decreasing tendency).[49]

Teupenhayn and Altmann (1984) did not reject the hypothesis for a sample of sentences from a French text (based on an F-test), and Hug (2004) showed a negative correlation between the tested unit lengths for almost all French newspaper articles when analysed individually (based on a linear correlation coefficient).

Kułacka (2009b) corroborated the law for excerpts from five out of seven Polish literary and scientific books (based on Spearman's correlation coefficient) and Teupenhayn and Altmann (1984) for Swedish, Hungarian, Slovak, Czech and Indonesian, each represented by a sample of sentences from an individual text (based on an F-test). The Czech was also tested by Benešová and Čech (2015), who confirmed the inverse proportionality between lengths of the sentence and clause in an essay (nevertheless, R^2 was below the standard of $R^2 \geq 0.90$, i.e. $R^2 = 0.8737$).

Berdicevskis (2021) tested 78 languages – each represented by the Universal Dependencies (UD) treebank (Universal Dependencies Treebanks 2.8.1, Zeman et al., 2021a) with more than 10,000 tokens or by their mixture in case a language had more than one treebank of such a size. The results showed a negative correlation between lengths of the units on this level for 68 languages (based on Spearman's rank correlation coefficient). As for the rest, none of the correlations was detected.

On the contrary, there are several studies which brought opposite results. When Hug (2004) tested the law on the whole sample of French newspaper articles, results revealed lengths of the sentence and the clause being correlated positively (based on a linear correlation coefficient). Roukk (2007) did not corroborate the hypothesis mentioned above for chapters of a German novel and their Russian translations, and a chapter of a Russian novel and its English translation (based on R^2).[50] The mean lengths of the clauses showed a zig-zag tendency rather than decreasing.[51] Even though Buk and Rovenchak (2008) did not interpret their test of the model reliability in detail, based on a curve visualising the fit between the model and their empirical data of a Ukrainian novel, the clause lengths showed an increasing tendency contradicting the law. However, it should be noted that their algorithm for clause determination struggled to cope with direct speeches.[52] Kułacka (2009b) did not corroborate the law for English and Polish excerpts from two books of different text types (based on Spearman's correlation coefficient).

Based on the summary of all the results, it appears that Ukrainian (Roukk, 2007) and Russian (Buk and Rovenchak, 2008) are the only languages where the menzerathian tendency was not observed. It is noteworthy that Berdicevskis (2021) also included Ukrainian and Russian in his study, and the results showed a negative correlation between the sentence and the clause

[49] The study also contains results for each text separately (available in Appendix, Heups, 1983, pp. 124-129).

[50] Roukk (2007, p. 605) also referred to her earlier works where she did not corroborate the law applied to speeches of children in Russian (the results should be published in Roukk, 2003a, 2003b).

[51] No model was applied.

[52] When comparing manual and automatic determination of the clause in the novel's first chapter, 19 % of clauses were miscounted (Buk and Rovenchak, 2008, p. 12).

lengths for both (although the methodology of all the three studies differs to a considerable extent).

Lastly, two studies tested the law while adopting a different approach to the analysis. First, Schwibbe (1984) fitted the models of the law only to mean sentence lengths and mean clause lengths calculated for each text of a different text type (essay, letter, fiction, scientific literature, news). His results showed the means being negatively correlated (based on an F-test). Second, Schwibbe (1989) tested letters and suicide notes, and texts written by two age groups while averaging the construct lengths that are usually discrete variables. In Schwibbe's view, the results were in accord with the law even though the author published only a graph showing the decreasing tendency of fitting curves.

2.1.1.2 The syntactic phrase as the constituent

The clause is not the only possible unit being the constituent to the sentence. Two studies used a syntactic phrase (or shortly phrase) measured in words and tested the hypothesis – the longer the sentence length measured in the number of phrases, the shorter the mean length of the phrases measured in words.

Mačutek, Čech and Milička (2017) approached the constituent in the realm of dependency grammar. The authors determined the phrase as a subtree directly hanging from a predicate of the main clause (=sentence)[53], and Tanaka-Ishii (2021) adopted the same approach. As for word determination, both the studies relied on an annotation scheme of language materials (Mačutek, Čech and Milička, 2017; Tanaka-Ishii, 2021).

Mačutek, Čech and Milička (2017) applied the law to the Prague Dependency Treebank 3.0 (Bejček et al., 2013; adjusted by the authors to some extent) and corroborated the hypothesis based on the coefficient of determination R^2 meeting the standard of $R^2 \geq 0.90$. Tanaka-Ishii (2021) opted for Universal Dependencies (UD) treebanks (Nivre et al., 2020, ver. 2.3) and the Penn Treebank (Marcus et al., 1994) converted to the dependency framework. In the case of UD, the author tested a sample consisting of 129 treebanks of 76 languages and each of the three largest UD treebanks – Czech, Russian and Japanese. The analysis of the mixed sample revealed that the phrase lengths monotonically decrease only for sentences of the length greater than two and lower than ten phrases (based on a decrease ratio)[54]. The decreasing tendency for a similar range was also revealed for the Czech, Russian and Japanese UD treebanks and even for the converted Penn Treebank.

[53] The authors called the construct a clause. However, language material under analysis distinguished only predicates of main clauses (Mačutek, Čech and Milička, 2017, p. 103). Therefore, we decided to introduce the study in this chapter.

[54] "The 'decrease ratio' indicates the average proportion of data points … for which the mean constituent size of x decreased as compared with that of $x - 1$" (Tanaka-Ishii, 2021, p. 4).

2.1.1.3 The clause and the syntactic phrase as the constituents

The last unit triplet is formed by combining the previous two direct constituents. Making the clause the direct constituent of the sentence and the syntactic phrase the direct constituent of the clause results in the last hypothesis – the longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in phrases.

The combination of these units has probably been analysed only in three studies (Sanada, 2016; Berdicevskis, 2021; Mačutek, Čech and Courtin, 2021). Sanada (2016) determined the number of clauses by the number of predicates (also allowing its non-verbal forms) and Mačutek, Čech and Courtin (2021) by the number of finite verbs. Berdicevskis (2021) used the annotation scheme of language material.

When it comes to the lowest unit, Sanada (2016) considered the phrase – called argument in her study – to be an element connected to the predicate. Berdicevskis (2021) followed the approach proposed by Mačutek, Čech and Milička (2017), i.e. the author treated the phrase as a sub-tree directly hanging from a predicate, even though he cast doubt on this operationalisation as well as Mačutek, Čech and Courtin (2021) who addressed its main drawbacks from several perspectives. Firstly, the previous approach resulted in constituent lengths higher than the short-term memory limit (e.g. roughly equal to 7 ± 2, as suggested by Miller, 1956). Secondly, the treatment of the predicate would either lead to the exclusion of sentences without phrases or to multiple inclusion of the predicate in each phrase. Lastly, the authors raised an objection that the original approach disregarded the linear property of a language. Taking these arguments into account, Mačutek, Čech and Courtin (2021) suggested a new unit – linear dependency segment (LDS) – defined as a group of words in a 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).[55]

Sanada (2016) tested the law on a set of Japanese sentences containing the verb 'meet', and her results corroborated its validity (based on the coefficient of determination R^2 meeting the standard of $R^2 \geq 0.90$). Berdicevskis (2021) yielded a negative correlation between the analysed unit lengths only in UD treebanks of 38 languages and a positive correlation in UD treebanks of five languages. In the case of 35, none of the correlations was proved (based on Spearman's rank correlation coefficient). Mačutek, Čech and Courtin (2021) analysed two Czech dependency treebanks converted to the Surface Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018). The authors corroborated the hypothesis only for one treebank and a sample which merges both (with $R^2 \geq 0.90$ being satisfied). Even though the fit of the second treebank was considerably lower (i.e. slightly above 0.61), the authors argued in favour of the apparent decreasing tendency of the LDS lengths.

[55] We are aware that the determination of the proposed unit does not solely rely on the dependency syntactic criterion but also takes the criterion of the word order into account. However, its position in the hierarchy of language units corresponds to a level between the clause and the word. Hence, we include it in chapters on the syntactic phrase.

2.1.2 The sentence in Chinese

Studies focusing on Chinese combined the sentence with the clause and the word. Hence, the hypothesis – the longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in words – was tested.

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[56] (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)[57], 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)[58] 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

[56] We remind the reader that the ellipsis strictly denotes the punctuation mark composed of three or six dots.

[57] Jin and Liu (2017) tested the approach on 1000 randomly selected sentences and found that the clausal segmentation based on the punctuation marks reached 95% accuracy. However, details about the test are not provided.

[58] 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).

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[59] out of 11 LCMC text collections. Jin and Liu (2017) showed corroborating results of four corpora of different text types – microblogs, news, prose and fiction (nonetheless, only the microblogs meet $R^2 \geq 0.90$). Chen (2018) and Chen and Liu (2019, 2022[60]) 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.[61] 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 as the construct

2.2.1 The clause across languages

Compared to the sentence level, the clause in the position of the construct has been studied to a considerably lesser extent and the clause determination varies across studies. The clause was determined based on the presence of a finite verb (Tuldava, 1995) or a predicate (Sanada, 2016), based on punctuation marks (Hug, 2004, who used a comma and dash as clause-final marks) or annotation of language material (Coloma, 2015, 2020; Berdicevskis, 2021 while applying minor modifications). As for the studies published by Benešová (2011), Andres and Benešová (2011, 2012) and Andres et al. (2012a), the construct under analysis was termed as sentence/clause (Andres and Benešová, 2011, 2012) or even syntactic construction (Benešová, 2011; Andres et al., 2012a) but always identified by its finite or infinitive verb functioning as a predicate (Benešová, 2011, p. 38; Andres et al., 2012a, p. 10). Based on this determination, it appears that rather the clause was analysed. Since the thesis considers the sentence to be a higher language unit which can include more than one predicate, we decided to introduce these studies within this chapter.

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

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

[61] 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.

2.2.1.1 The word as the constituent

When it comes to the direct and indirect constituents of the clause, studies mostly agree on the choice of the word to be the direct one but differ in the choice of a measurement unit for the word. Researchers opt for a syllable, a phoneme or a grapheme. Hence, the following hypothesis and its alternatives were tested – the longer the clause length measured in the number of words, the shorter the mean length of the words measured in syllables, phonemes or graphemes.

The word was determined as a sequence of graphemes between spaces (Hug, 2004; Benešová, 2011; Andres and Benešová, 2011; Andres et al., 2012a). Benešová (2011), Andres and Benešová (2011, 2012), and Andres et al. (2012a) also introduced an alternative approach to the word, which was regarded as a word form compounded from a carrier of a lexical meaning (e.g. noun) and a carrier of grammatical meaning (e.g. preposition, definite or indefinite article). Coloma (2015, 2020) and Berdicevskis (2021, with minor adjustments) used an annotation scheme of language material, and Tuldava (1995) did not address the determination of the word at all.

As for the lowest unit, the syllable was chosen by Tuldava (1995), Benešová (2011), Andres and Benešová (2011, 2012), and Andres et al. (2012a). However, details about its operationalisation are not included in these studies. Coloma (2015, 2020) opted for the phoneme and relied on phonetic transcription of language material. Hug (2004) and Berdicevskis (2021) combined the clause and the word with the grapheme. However, neither the phoneme nor the grapheme is usually perceived as the word direct constituent.

Most studies analysing the clause level show a certain degree of specificity in their approach. We start with those which do not methodologically diverge from mainstream works and end with studies that explicitly claim to apply the law while being methodologically on the borderline.

Similarly to the sentence level, Hug (2004) found, based on a linear correlation coefficient, that the clause length in words and the word length in graphemes were negatively correlated when French newspaper articles were tested separately (70 out of 103 articles). Otherwise, their mixed sample showed a positive correlation. The analysis of the identical triplet by Berdicevskis (2021) showed a negative correlation only in the case of UD treebanks of 12 languages. 29 of them were identified with a positive correlation. In the case of the rest (37), none of the correlations was confirmed (based on Spearman's rank correlation coefficient). However, as the author pointed out, the correlation coefficient is not informative for the treebanks showing a non-monotonic decrease.

Benešová (2011), Andres and Benešová (2011, 2012), and Andres et al. (2012a) deployed the law primarily for the identification of a language fractal. The authors applied the law to a Czech journalistic article (Benešová, 2011; Andres et al., 2012a) and the English poem 'The Raven' and its different translations into Czech, German (Benešová, 2011; Andres and Benešová, 2011) and Slovak languages (Andres and Benešová, 2012). From the perspective of the approach to the language fractal and follow-up fractal analyses, the authors concluded that selected samples corroborated the hypothesis (however, the question remains whether the hypothesis would be corroborated in these samples with regard to the standard of $R^2 \geq 0.90$, e.g. R^2 obtained from the journalistic article reached only the value of 0.65, Andres et al., 2012a, p. 28).

The last three studies differ in methodology to the largest extent. Coloma (2015, 2020) used transcriptions of a short fable in 100 languages provided by International Phonetic Association. Similarly to Schwibbe (1984), the author calculated the mean lengths of the clause and word per language transcription and fitted the law's models to these means. The fit expressed by the coefficient of determination R^2 was poor ($R^2 < 0.60$ when testing 50 languages in 2015 and $R^2 < 0.50$ when testing the second half of the languages in 2020). Tuldava (1995) primarily tested informational measures of dependency while using the law applied to an Estonian fiction text. However, the author did not draw any further conclusions about the relationship between the lengths of the clauses and words.

2.2.1.2 The syntactic phrase as the constituent

A syntactic phrase is the second alternative to the direct constituent of the clause. Even in this case, studies do not agree on the choice of the indirect constituent. The phrase is measured either in words or morphemes. Concerning these choices, the tested hypothesis was as follows – the longer the clause length measured in the number of phrases, the shorter the mean length of the phrases measured in words or morphemes.

As was already introduced above, Sanada (2016) determined the phrase as an element connected to a predicate and Berdicevskis (2021) followed the approach by Mačutek, Čech and Milička (2017), who determined the phrase as a whole subtree directly dependent on a predicate. It should be pointed out that Mačutek, Čech and Milička (2017) analysed only phrases belonging to predicates of main clauses, i.e. phrases were eventually the direct constituents of sentences.

Going to the lowest level, Sanada (2016) opted for a morpheme whose boundaries were identified by software and manual correction (although the question arises whether the morpheme is the direct phrasal constituent). Berdicevskis (2021) chose the word and relied on tokenisation of language material (with minor modifications).

As for the results, Sanada (2016) did not corroborate the law for a sample of Japanese sentences containing the verb 'meet' (the coefficient of determination R^2 did not exceed the value of 0.60). Berdicevskis (2021) revealed, based on Spearman's rank correlation coefficient, a negative correlation in UD treebanks of 58 languages and a positive correlation in UD treebanks of two languages. None of the correlations was detected for the rest (18).

2.2.2 Clause in Chinese

The clause in the position of the construct also occurred in studies focusing on Chinese. The clause was mostly combined with the word (being its direct constituent) and the Chinese character (being its indirect constituent), which resulted in the hypothesis – the longer the clause length measured in the number of words, the shorter the mean length of the words measured in Chinese characters.

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)[62], 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 erization (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.[63] When Hou et al. (2019b) refitted the data with the complete model, only the 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.[64] 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).[65]

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.[66]

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

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

[64] 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.

[65] 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.

[66] 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

2.3 The syntactic phrase as the construct

The syntactic phrase in the construct position was analysed only by Berdicevskis (2021), who chose the word and the grapheme as its direct and indirect constituents respectively and tested the hypothesis – the longer the phrase length measured in the number of words, the shorter the mean length of the words measured in graphemes. As mentioned above, Berdicevskis (2021) operationalised the phrase as a whole subtree which directly depends on a predicate and is measured in the number of words belonging to it (Mačutek, Čech and Milička, 2017). An annotation scheme of language material provided the word determination, and the word length was expressed as a sum of its graphemes (although the grapheme might not be the direct constituent of the word in all languages). Based on Spearman's rank correlation coefficient, a negative correlation between lengths of these units was identified only in UD treebanks of 11 languages, while a positive correlation in 22. None of the correlations was identifiable within the rest (45).

Berdicevskis (2021) also included Chinese in his analysis of the phrase level. The results showed that none of the correlations was confirmed.[67] 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 corroborated. Similarly, Sun and Shao (2021) added that the phrase might correspond to the clause.

2.4 The word as the construct

2.4.1 The word across languages

Several approaches to determining the word appeared among studies which tested the word in the position of the construct. The orthographical approach, i.e. identification of the word as a sequence of graphemes between two spaces, was followed by Alekseev (1998), Buk and Rovenchak (2007), Kelih (2008, 2010, 2012), Benešová (2011), Andres and Benešová (2011), Andres et al. (2012a), Benešová, Faltýnek and Zámečník (2015), Altmann and Gerlach (2016), and Torre, Dębowski and Hernández-Fernández (2021). Another approach was related to so-called zero-syllable words (i.e. particular non-vocalic words). As Wimmer et al. (2003, p. 105) addressed, these words should be either excluded from analysis or joined to words to which they relate. The exclusion of the zero-syllable words was applied by Grzybek (2000), Wimmer et

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).

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

al. (2003), Buk and Rovenchak (2007), Kraviarova and Zimmermann (2010) and Buk (2014). The second method, i.e. joining the zero-syllable words (e.g. prepositions, conjunctions, particles) to words that they either precede or follow, was adopted by Benešová (2011), Andres and Benešová (2011, 2012), Mačutek and Rovenchak (2011), Andres et al. (2012a), Mačutek, Chromý and Koščová (2018), Čech et al. (2020), Čech and Mačutek (2021) and Rujević et al. (2021). Similarly, Lehfeldt and Altmann (2002) concatenated the word and its neighbouring clitic(s) to one phonological word form (the determination of clitics in old Russian was based on a study by Zaliznjak, 1985). Benešová (2011), Andres and Benešová (2011, 2012), and Andres et al. (2012a) merged English and German articles with the following words (being called a compound analytic word form by the authors). There are also studies which re-analysed data published by earlier works (Altmann, 1980; Altmann and Schwibbe, 1989; Grzybek, 1999, 2000; Fenk, Fenk-Ozlon and Fenk, 2005; Mačutek and Rovenchak, 2011) or used available language material, i.e. words from dictionaries (Menzerath, 1954; Grzybek, 2000; Köhler 2002; Dinu and Dinu, 2009) or a tokenised corpus (Araujo, Benevides and Pereira, 2020). Mikros and Milička (2014) and Galieva (2021) developed a script for text processing but did not directly address tokenisation. Lastly, some studies provided results but not details how the word was operationalized (Altmann and Schwibbe, 1989; Fenk and Fenk-Oczlon, 1993; Hřebíček, 1995; Fenk, Fenk-Oczlon and Fenk, 2005; Milička, 2014; Best and Rottmann, 2017). Since word recognition based on the orthographical criterion is commonly used in quantitative linguistics, it can be expected that the authors adopted this pragmatic approach.

2.4.1.1 The syllable as the constituent

The syllable as the direct constituent of the word prevails. As for the word indirect constituent, a phoneme (or sound) and grapheme were chosen, while the former was usually preferred. The combination of these units led to the following hypothesis – the longer the word length measured in the number of syllables, the shorter the mean length of the syllables measured in phonemes or graphemes.

Regarding the determination of the syllable, the sonority of syllabic elements and the sonority sequencing principle were employed. According to the principle, the highest degree of the sonority is assigned to a syllabic nucleus (vowel or syllabic consonant) representing its sonority peak (e.g. Hall, 2006, p. 330). Hence, the number of peaks equals the number of syllables in the word. The sonority sequencing principle was deployed either for the automatic segmentation of the words into syllables (Rujević et al., 2021, who combined it with the maximum onset principle for consonants in an intervocalic position; Torre, Dębowski and Hernández-Fernández, 2021), or just for the determination of the number of syllables in the word (Menzerath, 1954; Kelih, 2008, 2010, 2012; Mačutek and Rovenchak, 2011; Mikros and Milička, 2014; Mačutek, Chromý and Koščová, 2018; Čech and Mačutek, 2021). Alekseev (1998) relied on graphemics in connection with syllable borders but did not provide further details. Dinu and Dinu (2009) and Araujo, Benevides and Pereira (2020) used an annotation of language material, while Altmann and Gerlach (2016) relied on the Moby Hyphenation List (Ward, 2002). Some studies re-analysed previously published data (Altmann, 1980; Altmann and Schwibbe, 1989; Grzybek, 1999, 2000; Fenk, Fenk-Oczlon and Fenk, 2005; Mačutek and Rovenchak, 2011). Last but not least, there are a number of studies which lack information about the syllable operationalization (Altmann and Schwibbe, 1989; Fenk and Fenk-Oczlon, 1993; Hřebíček, 1995; Grzybek, 2000; Köhler, 2002; Lehfeldt and Altmann, 2002; Wimmer et al., 2003; Fenk, Fenk-Oczlon and Fenk, 2005; Buk and Rovenchak, 2007; Kraviarova and Zimmermann, 2010; Benešová, 2011; Andres and Benešová, 2011, 2012; Andres et al., 2012a; Buk, 2014; Milička, 2014; Benešová, Faltýnek and Zámečník, 2015; Best and Rottmann, 2017; Čech et al., 2020; Galieva, 2021).

As for the indirect constituents of the word, we introduce their determination according to units for which researchers explicitly opted. Let us start with the phoneme. Some studies used graphemes and converted them into phonemes based on rules specific to a language under analysis (Menzerath, 1954, who termed the unit as the sound; Köhler, 2002; Mačutek and Rovenchak, 2011; Mikros and Milička, 2014; Rujević et al., 2021; probably also in Lehfeldt and Altmann, 2002[68]). Mačutek, Chromý and Koščová (2018) directly counted graphemes due to their close correspondence to the phonemes in Czech. Galieva (2021) only referred to the phoneme as one symbol in Tatar without specifying its operationalisation. Altmann and Gerlach (2016) used The CMU Pronouncing Dictionary version 0.7b (Carnegie Mellon University, n.d.) for the phoneme determination and Araujo, Benevides and Pereira (2020) relied on a phonetic transcription of language material. Some data were just re-analysed and the indirect constituent called either as the phoneme (Altmann, 1980; Altmann and Schwibbe, 1989; Fenk, Fenk-Oczlon and Fenk, 2005; Mačutek and Rovenchak, 2011) or sound (Grzybek, 1999, 2000). Lastly, authors which also used the inventory of phonemes or (speech) sounds of a language under analysis are as follows Altmann and Schwibbe (1989), Fenk and Fenk-Oczlon (1993), Hřebíček (1995), Wimmer et al. (2003), Fenk, Fenk-Oczlon and Fenk (2005), Kraviarova and Zimmermann (2010), Benešová (2011), Andres and Benešová (2011, 2012), Andres et al. (2012a), Buk (2014), Milička (2014), Best and Rottmann (2017), Čech et al. (2020) and Čech and Mačutek (2021).

The lowest unit was explicitly termed as the grapheme by Alekseev (1998), Grzybek (2000), Kelih (2008, 2010, 2012), Benešová, Faltýnek and Zámečník (2015) and Torre, Dębowski and Hernández-Fernández (2021). However, as Kelih (2012, p. 205) pointed out when analysing Slovene, the grapheme closely corresponds to the phoneme in Slavic languages (cf. Mačutek, Chromý and Koščová, 2018, who called the lowest unit the phoneme when analysing Czech).

Lastly, Buk and Rovenchak (2007) and Dinu and Dinu (2009) analysed in their studies both – phoneme and grapheme. However, when presenting the results of the law's application, the former study did not specify the choice, and the latter study termed the indirect constituent as the phoneme but used the term letter for quantitative and descriptive properties of language material.

As for the summary of results, we evaluate language material under analysis concerning word tokens and word types[69] due to their impact on the results (see Chapter 1.4) and follow

[68] The authors only mentioned phonological interpretation of letter sequences without a further specification (Lehfeldt and Altmann, 2002, p. 38).

[69] We remind the reader that we use the notion 'types' not only for basic forms of words, i.e. lemmas, but also for different word forms of a lemma (Taylor, 2015, pp. 2-3).

interpretation published by authors. Firstly, studies on the triplet of the word, syllable and phoneme are summarised.

When starting with the Czech language, Milička (2014) primarily tested different models of the law on word tokens from a novel. The coefficient of determination R^2 reached a value higher than 0.90 at least in the case of a formula derived by the author.[70] Mačutek, Chromý and Koščová (2018) did not reject the hypothesis for word tokens from most of the analysed interviews and Čech et al. (2020) for words from individual texts and their mixed sample (however, whether the tokens or the types were analysed is not specified). Both the studies followed the standard of $R^2 \geq 0.90$. Finally, Čech and Mačutek (2021) tested the types, and their results showed the corroboration of the hypothesis for most of the poetic and prosaic samples while also following $R^2 \geq 0.90$ (only R^2 of two poems out of 13 reached lower values, one of which was slightly below the standard, i.e. $R^2 = 0.883$).

Wimmer et al. (2003) and Kraviarova and Zimmermann (2010) yielded corroborating results for the Slovak language represented by word tokens from a poem (Wimmer et al., 2003, with R^2 exceeding 0.90) and word tokens from separate text excerpts (Kraviarova and Zimmermann, 2010). However, the authors of the latter study did not provide the goodness-of-fit between a model and data, and only word lengths greater than one syllable showed the menzerathian decreasing tendency.

As regards the Serbo-Croatian language, Altmann and Schwibbe (1989) and Grzybek (1999, 2000) corroborated the hypothesis for word types from a dictionary published by Gajić (1950), the former study based on an F-test, whereas the latter study based on R^2 (being in accord with $R^2 \geq 0.90$). Rujević et al. (2021) tested word tokens from Serbian and Croatian translations of Russian chapters. The excellent fit (as evaluated by the authors, R^2 is not available) was achieved only when the authors fitted their alternative model with four parameters to the data.[71] As mentioned in Chapter 1.3, the more parameters a model has, the better results might be obtained, however, at the cost of lower interpretability of additional parameters. Hence, "models with more than three parameters … are seldom useful in linguistics" (Köhler, 2012, p. 53).

As for the Russian language, Lehfeldt and Altmann (2002; also in Lehfeldt, 2007) did not reject the hypothesis for words from text samples representing old Russian after the fall of jers[72] (R^2 exceeded the value of 0.90). The authors did not specify whether the tokens or the types were analysed. However, their approach implies the tokens. Rujević et al. (2021) tested Russian on word tokens from chapters of a novel and the hypothesis was corroborated only when the authors fitted their model to the data (the goodness-of-fit met the standard $R^2 \geq 0.90$).[73]

[70] The alternative model is $L_{n-1} = a_n + \frac{b_n}{L_n} + \frac{c_n \min(1, L_n - 1)}{L_n}$, where L_n is the construct length of a level n, L_{n-1} is the mean length of its constituents, a_n, b_n and c_n are parameters (Milička, 2014).

[71] The alternative model is $y(x) = ax^{b+c\log x}e^{-dx}$, where $y(x)$ is the mean length of the constituent of a given construct x, and a, b, c and d are parameters (Rujević et al., 2021).

[72] I.e. vowels ь and ъ which were reduced in Russian (Lehfeldt and Altmann, 2002, pp. 39-41).

[73] The alternative model is $y(x) = ax^{b+c\log x}e^{-dx}$, where $y(x)$ is the mean length of the constituent of a given construct x, and a, b, c and d are parameters (Rujević et al., 2021).

Regarding studies on Ukrainian, Mačutek and Rovenchak (2011) corroborated the hypothesis when applying the law to canonical word form types from different text types and their mixture.[74] The authors followed the standard $R^2 \geq 0.90$ but used a modified formula.[75] Buk (2014) did not reject the hypothesis for word types from a novel, however, $R^2 \geq 0.90$ was also obtained by fitting an alternative model to the data.[76] Contrary to these two studies, Rujević et al. (2021) tested Ukrainian on word tokens from translated chapters of a Russian novel and only the formula derived by the authors yielded an excellent fit (R^2 is not available).[77]

Altmann (1980) and Altmann and Schwibbe (1989) re-analysed English word types from data published by Roberts (1965). Even though the studies differ in word lengths, both corroborated the hypothesis (the latter study based on an F-test with the apparent menzerathian decreasing tendency visualised). Altmann and Gerlach (2016) did not reject the law for English word types from a sample consisting of a book.[78] Although the authors primarily tested different models including Altmann's complete formula by a likelihood method, the fit between the models and the data showed the apparent menzerathian decreasing tendency.

Altmann and Schwibbe (1989) and Fenk, Fenk-Oczlon and Fenk (2005) re-analysed German word types from a dictionary based on which Menzerath (1954) came to his conclusion. The studies did not reject the hypothesis. The former used an F-test and showed a fitting curve following the menzerathian decreasing trend. The latter used the coefficient of determination R^2 and showed a fit being in accord with $R^2 \geq 0.90$. The hypothesis was also not rejected by Best and Rottmann (2017), who tested both – word tokens and word types – from a German prose text. The types satisfied $R^2 \geq 0.90$ while R^2 for the tokens was slightly below the standard, i.e. $R^2 = 0.88$. The data was, however, fitted by an alternative formula.[79]

Araujo, Benevides and Pereira (2020) corroborated the law for word types from a Brazilian Portuguese corpus (based on the menzerathian decreasing trend while using a logarithmic transformation).[80] In the case of Italian, Altmann and Schwibbe (1989) and Fenk,

[74] The phonemes are reduced only to vowels and consonants in these words (Mačutek and Rovenchak, 2011, p. 136).

[75] The alternative model is $S_P(W_S) = aW_S^b + 1$, where S_P is the mean length of syllables measured in the number of phonemes, W_S is the word length measured in the number of syllables, a and b are parameters and 1 is a constant added with respect to syllables having at least one phoneme (Mačutek and Rovenchak, 2011).

[76] The alternative model is $L(s) = L_\infty + Bs^c$, where L is the mean syllable length measured in phonemes, s is the word length measured in syllables, L_∞ is the mean syllable length in a hypothetically infinite word, B and c are parameters (Buk, 2014).

[77] The alternative model is $y(x) = ax^{b+c\log x}e^{-dx}$, where $y(x)$ is the mean length of the constituent of a given construct x, and a, b, c and d are parameters (Rujević et al., 2021).

[78] Altmann and Gerlach (2016) also tested English word types from an English Wikipedia. However, we are not able to evaluate the law's corroboration based only on the results of the likelihood analysis presented in the study.

[79] The alternative model is $y = ax^b e^{(\frac{c}{x}+dx)}$, where y is the mean constituent lengths, x is the construct length, a, b, c and d are parameters (Best and Rottmann, 2017).

[80] The results and the follow-up discussion in Araujo, Benevides and Pereira (2020) imply that the decreasing tendency concerns the types. Despite the decrease in the syllable lengths, the coefficient of

Fenk-Oczlon and Fenk (2005) re-analysed word types from a dictionary published by Rettweiler (1950; and re-published by Menzerath, 1954). While the former authors did not reject the hypothesis based on an F-test and the visualised menzerathian decreasing tendency, the latter authors achieved the standard of $R^2 \geq 0.90$ only when they used their polynomial model.[81] Mikros and Milička (2014) confirmed the monotonical decreasing tendency for word types and not word tokens when analysing a Greek corpus (the data was not fitted with any model). However, when Rujević et al. (2021) fitted these Greek word tokens with their formula, the result showed an excellent fit.[82] Köhler (2002) analysed Hungarian word types from a dictionary and the hypothesis was not rejected (the standard of $R^2 \geq 0.90$ was reached). While taking fluctuations in constituent lengths into account, Galieva (2021) concluded that Tatar word tokens from poems and prosaic texts showed the general menzerathian tendency and a reasonably good fit. Nonetheless, only three out of six samples would meet $R^2 \geq 0.90$. Hřebíček (1995) corroborated the hypothesis for words from a Turkish text, but the choice of tokens or types was not specified ($R^2 \geq 0.90$ was satisfied). Finally, there are corroborating results for Indonesian word types which were analysed by Altmann and Schwibbe (1989, based on an F-test with the apparent menzerathian decreasing tendency)[83] and Indonesian canonical word form types, which were initially published by Altmann et al. (2002) but re-analysed by Mačutek and Rovenchak (2011). The authors yielded a satisfactory fit (i.e. above $R^2 \geq 0.90$) while using their alternative formula.[84]

The results which did not corroborate the law according to authors were yielded for Czech by Mačutek, Chromý and Koščová (2018) when testing word tokens from two interviews and by Čech and Mačutek (2021) when testing word types from a poem. Lehfeldt and Altmann (2002; also in Lehfeldt, 2007) also rejected the hypothesis. The authors applied the law to words from a text excerpt representing old Russian before the reduction of the jers (the methodology implies the analysis of the tokens). On the one hand, R^2 of both Altmann's – truncated and complete – models did not reach or approximated $R^2 \geq 0.90$, on the other hand, a model suggested by the authors accorded with the standard).[85] Buk (2014) rejected the hypothesis for word tokens from a Ukrainian novel even if the author additionally tested direct and author's speeches separately. Mikros and Milička (2014) showed that the monotonic decreasing

determination R^2 reached low values, which the authors explained by applied models determined for all data in a sample and not only for commonly used averages (Araujo, Benevides and Pereira, 2020, p. 39).

[81] The alternative model is not provided.

[82] The alternative model is $y(x) = ax^{b+c\log x}e^{-dx}$, where $y(x)$ is the mean length of the constituent of a given construct x, and a, b, c and d are parameters (Rujević et al., 2021).

[83] The same data were probably re-analysed by Fenk, Fenk-Oczlon and Fenk (2005), who showed the fit reaching the standard of $R^2 \geq 0.90$. However, the authors referred to Menzerath (1954), although Menzerath (1954) did not publish any data on Indonesian.

[84] The alternative model is $S_P(W_S) = aW_S^b + 1$, where S_P is the mean length of syllables measured in the number of phonemes, W_S is the word length measured in the number of syllables, a and b are parameters and 1 is a constant added with respect to syllables having at least one phoneme (Mačutek and Rovenchak, 2011).

[85] The alternative model is $y = Kx^{-b}e^{-ax}e^{c/x}$, where y is the mean constituent length, x is the construct length, K, a, b and c are parameters (Lehfeldt and Altmann, 2002).

tendency is violated when Greek word tokens from a text and a corpus are analysed (the data was not fitted).

Lastly, there are studies which are specific in their approach. As mentioned in Chapter 2.2.1.1, Benešová (2011), Andres and Benešová (2011, 2012), and Andres et al. (2012a) used the law primarily for the fractal analysis. The authors tested a Czech journalistic article and the English poem 'The Raven' and its different translations into Czech, German and Slovak. It can be concluded based on their approach that some samples corroborated the law (even though the journalistic text, for example, did not reach $R^2 \geq 0.90$, Andres et al., 2012a, p. 28). Fenk and Fenk-Ozclon (1993) and Fenk, Fenk-Oczlon and Fenk (2005) firstly calculated the mean word lengths (in syllables) and the mean syllable lengths (in phonemes) for each language under analysis and then tested the menzerathian relationship on these means. Both studies yielded similar results – values of R^2 were very low. However, when Fenk, Fenk-Oczlon and Fenk (2005) followed the standard methodology, their sample of three different languages satisfied $R^2 \geq 0.90$.

Let us summarise results from studies that opted for the grapheme as the indirect constituent. Kelih (2008) corroborated the hypothesis for Czech word types from a translation of a Russian text (results met the standard of $R^2 \geq 0.90$) and Benešová, Faltýnek and Zámečník (2015) for Czech word tokens from a dialogue transcription. As for the latter study, the authors considered the tendency being in accord with the law, although not fully satisfied. Nonetheless, a value of R^2 was considerably below the standard, i.e. $R^2 = 0.6253$. The hypothesis was not rejected by Kelih (2008) for Macedonian word types from a translated Russian text and by Kelih (2010) for Serbian word types from different text types and their corpus. Both the studies showed R^2 in agreement with $R^2 \geq 0.90$. The same results (i.e. R^2 being above 0.90) were achieved in the case of Slovenian word types from a dictionary (Grzybek, 2000), translated Russian text (Kelih, 2008) and different text types and their mixture (Kelih, 2012). As for the Russian language, Kelih (2008) did not reject the hypothesis for word types from a novel (with respect to $R^2 \geq 0.90$) and Alekseev (1998) showed the decreasing tendency for Russian word types from a sample of letters. Russian word tokens tested by Alekseev (1998) on individual text types and a whole corpus violated this tendency by reaching the maximum of syllable lengths with 2-syllable or even 3-syllable words. Lastly, Torre, Dębowski and Hernández-Fernández (2021) applied the law to word tokens from individual samples of 21 languages. The authors excluded monosyllabic words from the analysis and evaluated the results based on Spearman's rank correlation coefficient. Approximately half of the languages followed the monotonically decreasing tendency, whereas almost all of them corroborated the second regime of the law (as discussed in Chapter 1.3).

Studies which lack precise information about the indirect constituent under analysis showed the menzerathian decreasing tendency of fitting curves – Buk and Rovenchak (2007) while applying an alternative formula to Ukrainian word types from a novel[86], and Dinu and Dinu (2009) in the case of Romanian word types from a dictionary.

[86] The alternative model is $M = M_\infty + Bs^c$, where M is the mean syllable length, s is the word length, M_∞ is the mean syllable length in a hypothetically infinite word, B and c are parameters (Buk and Rovenchak, 2007).

Finally, we briefly outline studies which analysed combinations of units beyond the scope of this thesis while keeping the word as the construct. Rovenchak (2015) brought results when applying the law to the word measured in syllables and the syllable measured in moras. Gerlach (1982)[87], Krott (1996), Hřebíček (1995, 2002a), Milička (2014), and Pelegrinová, Mačutek and Čech (2021) measured the word directly in morphemes and indirectly in phonemes (or sounds). Some studies also tested the word measured in morphemes but opted for the grapheme as its sub-constituent, e.g. Krott (1996), Polikarpov (2000)[88], Benešová, Faltýnek and Zámečník (2015) and Stave et al. (2020). Researchers also analysed the word level while measuring its constituents (syllables or phonemes) in time units, e.g. Altmann (1980) [89], Hernández-Fernández et al. (2019) and Torre et al. (2019). Lastly, there are studies on the relationship between the word length measured either in syllables or graphemes and the mean number of word meaning(s) while fitting data with the law's models, e.g. Altmann, Beöthy and Best (1982), Rothe (1983), Fickermann, Markner-Jäger and Rothe (1984), Sambor (1984), and Schwibbe (1984). By the optics of the menzerathian relationship, the question arises whether the number of the meanings can be considered the sub-constituent to the word and, therefore, its measurement unit.

2.4.2 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). Concerning this variability, the following hypotheses were tested – 1) the longer the word length measured in the number of syllables, the shorter the mean length of the syllables measured in phonemes or graphemes, and 2) the longer the word length measured in the number of Chinese characters, the shorter the mean length of the Chinese characters measured in 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[90].

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

[87] Later re-analysed by Altmann and Schwibbe (1989), Milička (2014) and Best and Rottmann (2017).

[88] Later re-analysed by Milička (2014).

[89] Also in Geršić and Altmann (1980) and Altmann and Schwibbe (1989).

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

was used (without a reference). The grapheme was determined as a Latin letter of pinyin transcription.[91] The study analysed word tokens from a corpus of dialogic text and did not corroborate the hypothesis either for the phoneme or the grapheme.[92]

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 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 Character as the construct

The last language unit being the construct tested within this thesis is a basic unit of Chinese and Japanese writing systems – the character – being measured directly in its components and indirectly in its strokes. Hence, the final hypothesis is as follows – the longer the character measured in the number of components, the shorter the mean number of the components measured in strokes.

In general, the character always occupies a graphic field of the same size without regard to its complexity. When the language is considered, the writing systems differ. The Chinese script is rather homogenous – either in its simplified or traditional form. The Japanese script combines three different types of characters, i.e. logographic Chinese characters known as kanji and syllabary characters known as kana (hiragana and katakana). The Chinese characters have been analysed so far by Bohn (1998, 2002), Motalová et al. (2013), Motalová and Matoušková (2014),

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

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

Matoušková and Motalová (2015) and Matoušková (2016). Prün (1994) tested the kanji characters, and Benešová and Birjukov (2015) and Birjukov (2016) the Japanese script in its complex form.

The component is generally considered a structural unit smaller than the character but greater than the stroke. As for its precise determination, Prün (1994) opted for a list of components of kanji characters compiled by Stalph (1989), which was also used by Bohn (1998, 2002) with slight modifications. 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; Benešová and Birjukov, 2015; Matoušková and Motalová, 2015; Birjukov, 2016; Matoušková; 2016). Regarding the strokes, each character in both languages has its immutable inventory.

In the case of types, the hypothesis was corroborated for kanji characters from a list of regular – jôyôkanji – characters (Prün, 1994) and 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 from one translation of the poem 'The Raven' (Matoušková and Motalová, 2015) and studies by Benešová and Birjukov (2015) and Birjukov (2016), who analysed individual Japanese texts including all the three types of the characters (kanji, hiragana and katakana). All the studies tested tokens.

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., 2016; 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., 2021b). 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).[93] 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.

The Chinese-HK UD treebank (Wong et al., 2017)[94] was manually annotated using the UD framework. It contains 1004 sentences from two sources which considerably differ in their properties. The first source (650 sentences) combines subtitles of three short movies, which mainly include informal utterances of various speakers composed of short sentences. The second source (354 sentences) is an excerpt from proceedings of a presidential election during a legislative council meeting[95]. Utterances of speakers are rather formal and sentences are longer (the mean sentence length in words is 12.18 in the proceedings while 5.88 in the subtitles, Poiret et al., 2021, p. 23). From the perspective of the heterogeneity having an impact on results, as discussed in Chapter 1.4, the subtitles mix different movies, or in other words, different contexts. Hence, the degree of their heterogeneity is higher. Moreover, the short sentences might prevent the law from coming into force, as pointed out, for example, by Kułacka (2009b), Jin and Liu (2017) and Hou et al. (2017), who tested samples of conversational nature. For this reason, we decided to split the treebank and to analyse only the proceedings, labelled as HK-P, which does not distort the material homogeneity as much as the subtitles.

The Chinese Parallel Universal Dependency treebank (Zeman et al., 2017)[96], labelled as PUD, was automatically transformed into the UD framework. The treebank consists of 1000 sentences randomly selected from news and Wikipedia. The sentences were originally collected from different language sources (most of the sentences – 750 – were written in English) and subsequently translated by professional translators into target languages using only English versions. We analyse the treebank as a whole (labelled as PUD). However, due to the usage of

[93] 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.

[94] Information is also available in the UD online guideline (UD Chinese HK, 2021) and on Github (UD_Chinese-HK, 2021).

[95] The meeting of the Legislative council of the Hong Kong Special Administrative Region of the People's Republic of China (HKSAR) on 12th October 2016 (Poiret et al., 2021, p. 23).

[96] Information is also available in the UD online guideline (UD Chinese PUD, 2021) and on Github (UD_Chinese-PUD, 2021).

the two different sources, which might influence the degree of heterogeneity, we also perform the analyses on data from the news (500 sentences labelled as PUD-N) and Wikipedia (500 sentences labelled as PUD-W) separately.

The last UD treebank, i.e. Chinese GSDSimp treebank (UD Chinese GSDSimp, 2021)[97], labelled as GSD, was also automatically converted into the UD framework and includes 3997 sentences collected from Wikipedia. Contrary to the PUD treebank, only one source was used to collect the sentences. Hence, the whole treebank is analysed.

Last but not least, the additional sample of The Lancaster Corpus of Mandarin Chinese (McEnery, Xiao and Mo, 2003), labelled as LCMC, contains 45,590 sentences collected from texts of 15 different text types written in mainland China. All texts are segmented into paragraphs, sentences and words carrying part-of-speech annotation. The corpus does not annotate clauses and dependency relations which are crucial for determining higher linguistic levels tested in the thesis. Hence, we exploit the sample only on the word and character level.

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

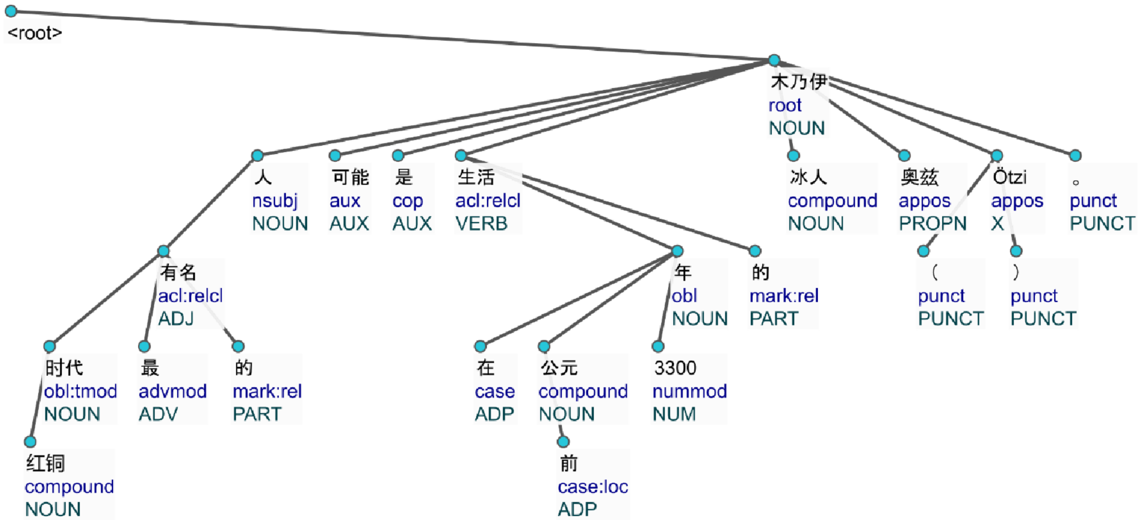
The chapter describes the determination and operationalisation of language units we chose to analyse with respect to the Chinese language and the assumption that the menzerathian relationship between the construct and the constituent lengths occurs when neighbouring units are tested.

3.2.1 The sentence

The sentence is represented in UD as a tree (see Figure 3), 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; Syntax: General Principles,

[97] Information is also available on Github (UD_Chinese-GSDSimp, 2021).

2021). Only one word is promoted to be a head of the whole sentence – called root[98] – while the rest of the words directly or indirectly – through other words – depends on it.



红铜时代最有名的人可能是生活在公元前 3300 年的木乃伊冰人奥兹（Ötzi）。

Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).

'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'

Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author.

Figure 3. The example of a sentence in the form of a UD tree (sentence ID w02008038, PUD treebank).

As for the governance of a dependency relation between two words, the priority is given to content words while function words directly depend on them (Nivre et al., 2020, pp. 4035-4036; de Marneffe et al., 2021, p. 257; The Primacy of Content Words, 2021). Hence, nodes in dependency trees are arranged rather horizontally than vertically. Or in other words, the trees grow rather into the breadth than the depth, which flattens syntactic structures and impacts the lengths of given linguistic levels (e.g. phrases). This choice, however, comes under criticism. The main objection is mixing semantic and syntactic criteria – "positioning content words over function words is a semantic criterion, but the actual annotation choices are expressed in terms of syntactic category, a syntactic criterion" (Osborne and Gerdes, 2019, p. 10).[99]

To decompose the structure of the sentence into its smaller parts, i.e. clauses, their heads, i.e. predicates, are taken into account. If the sentence consists only of one predicate (i.e. a root), it is categorised as a simple sentence (or in terms of UD as a simple clause, e.g. de Marneffe et al., 2021, pp. 272-276; Simple Clauses, 2021). If two or more predicates are

[98] In the UD perspective, the root is only a notional node (labelled as <root> in Figure 3) which a sentential head (木乃伊, *mùnǎiyī*, 'mummy') depends on via the `root` dependency relation (Syntax: General Principles, 2021). However, we call the head of a whole sentential structure the root for easier reference.

[99] E.g. There is an alternative to UD, which builds solely on syntactic criteria, called Surface Syntactic Universal Dependencies (SUD). SUD represents a surface-syntactic annotation scheme for dependency treebanks which prioritizes function words over content words, contrary to UD (Gerdes et al., 2018).

identified (one being the root while the rest being heads of other simple clauses), the sentence is classified as a complex sentence (or in terms of UD as a complex construction or complex clause, e.g. de Marneffe et al., 2021, pp. 276-279; Complex Clauses, 2021).

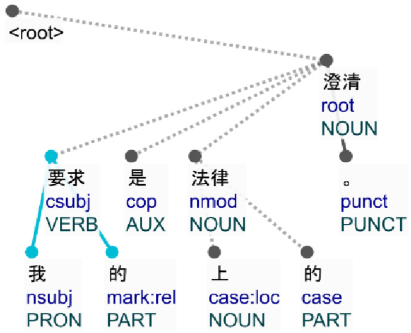
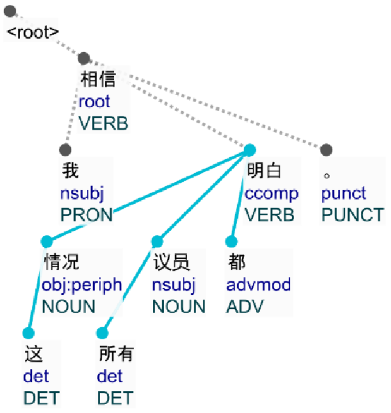
The clauses in the complex sentence are interconnected either through coordination or subordination. Coordination (de Marneffe et al., 2021, pp. 276-277; Coordination, 2021) occurs when two or more clauses of the same level are identified (with or without conjunction between them). Despite their symmetric relation and heads being of the same level, the dependency tree structure does not allow them to be treated equally. The first predicate governs the whole coordinate structure and predicates of other clauses depend on it via the UD conjunct relation (`conj`) while respecting their linear order. Subordination (de Marneffe et al., 2021, pp. 27-278; Subordination, 2021) emerges between two clauses of different levels. From the view of the dependency tree, a predicate of a subordinate clause directly depends on its governor which belongs to a higher clause and which the subordinate clause develops.

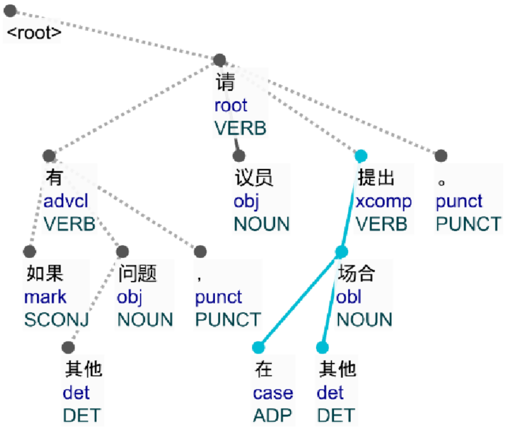
3.2.2 The clause

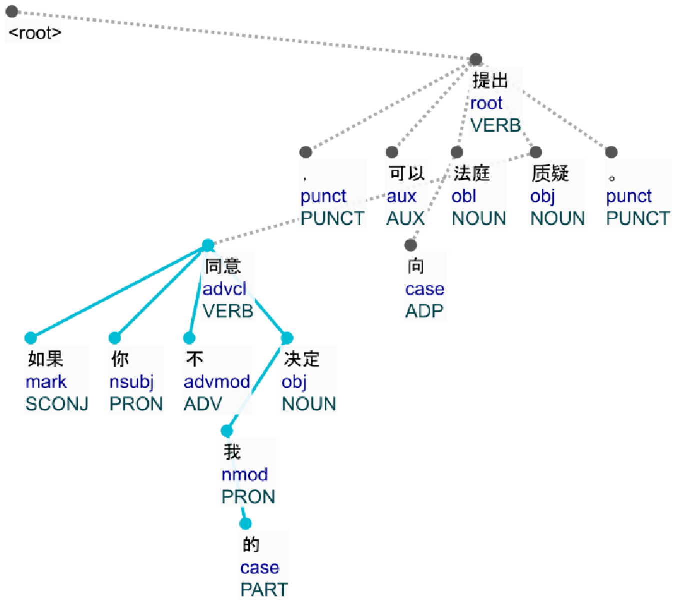
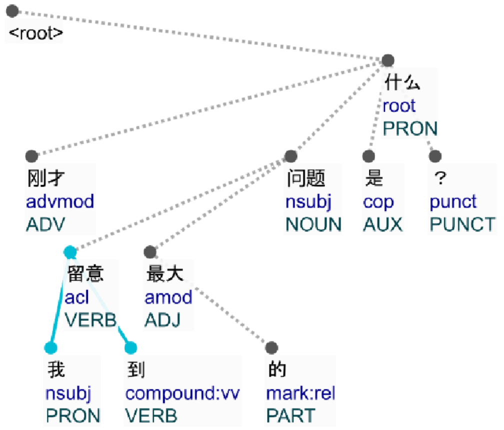
The simple clause consists of a head, i.e. verbal or non-verbal predicate, and its directly or indirectly dependent words (if any). The simple clause can correspond to a sentence with only one predicate (a root) and, consequently, can be represented by a whole tree. Otherwise, it is a subtree corresponding 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, 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`).

The annotation of UD treebanks is crucial to our analysis since determining the clause in Chinese encounters numerous difficulties (as pointed out, for example, by Hou et al., 2017; Jin and Liu, 2017; Xu and He, 2018). Clauses are commonly determined based on the presence of predicates expressed by finite verbs (applied, for example, by Köhler, 1982; Heups, 1983; Teupenhayn and Altmann, 1984; Roukk, 2007; Benešová and Čech, 2015; Xu and He, 2018; Mačutek, Čech and Courtin, 2021; Jiang and Jiang, 2022). However, this approach is not applicable to the Chinese language. The verbs cannot be inflected (only joined with aspect markers, e.g. Li, 2016, p. 81), and they are not the only category which functions as the predicate in Chinese. Adjectives also typically occupy the predicate's position (Huang, Jin and Shi, p. 276), and other non-verbal categories are allowed too, e.g. prepositions (Li, 2016, p. 88) or nouns (Shi, 2016, p. 249). For this reason, our clause determination is not conditioned by any additional rule and we rely entirely on the UD annotation for the clausal dependency relations described above. The following overview (Table 3) provides descriptions and examples to illustrate how the subordinate clauses are determined for Chinese in UD (de Marneffe et al., 2021, p. 266 and pp. 277-280; Universal Dependency Relations, 2021; Dependencies, 2021).

Table 3. Overview of subordinate clauses in UD.

Clause	Description & Example	UD label
Clausal subject	The clause functions as an active or a passive subject of a predicate. (See csubj: clausal subject, 2021a; csubj: clausal subject, 2021b)  <root> 澄清 root NOUN 要求 csubj VERB 是 cop AUX 法律 nmod NOUN 。 punct PUNCT 我 nsubj PRON 的 mark:rel PART 上 case:loc NOUN 的 case PART 我要求的是法律上的澄清。 *Wǒ yāoqiú de shì fǎlǜ shàng de chéngqīng.* ‘I ask for a legal clarification.’ (sentence ID 742, HK-P treebank)	csubj csubj:pass
Clausal complement	The clause functions as an object. It is not obligatory for a subject of the clausal complement to refer to any argument within its governing clause. In case of its omission, the subject is known and pragmatically understood. The relation is also applied if the clausal complement: 1) follows a verb + 得 (*de*, particle) together with its subject, 2) is in the position of a copula’s argument (是, *shì*, ‘to be’), 3) follows the head 是 (*shì*, ‘to be’) in construction 是…的(*de*, grammatical particle). (See ccomp: clausal complement, 2021a; ccomp: clausal complement, 2021b)  <root> 相信 root VERB 我 nsubj PRON 明白 ccomp VERB 。 punct PUNCT 情况 obj:periph NOUN 议员 nsubj NOUN 都 advmod ADV 这 det DET 所有 det DET 这情况我相信所有议员都明白。 *Zhè qíngkuàng wǒ xiāngxìn suǒyǒu yìyuán dōu míngbai.* ‘I believe that all members of the legislative body understand this situation.’ (sentence ID 811, HK-P treebank)	ccomp

Open clausal complement	The clause also functions as the object. Unlike ccomp, the open clausal complement has a subject which is obligated to unambiguously refer to an argument in a governing clause, i.e. subject or direct object. The relation is also applied to secondary predicate, optional and obligatory resultatives, obligatory depictives, the construction verb + 得 (*de*, particle) followed by the open clausal complement without the subject or a particular predicative adjective. (See xcomp: open clausal complement, 2021a; xcomp: open clausal complement , 2021b)  <root> 请 root VERB 有 advcl VERB 议员 obj NOUN 提出 xcomp VERB 。 punct PUNCT 如果 mark SCONJ 问题 obj NOUN ， punct PUNCT 场合 obl NOUN 其他 det DET 在 case ADP 其他 det DET 如果有其他问题，请议员在其他场合提出。 *Rúguǒ yǒu qítā wèntí, qǐng yìyuán zài qítā chǎnghé tíchū.* 'If there are other questions, I ask members of the legislative body to raise them on another occasion.' (sentence ID 709, HK-P treebank)	xcomp
Adverbial clause modifier	The clause represents a – temporal, conditional, purpose, consequence – adjunct which modifies a predicate or modifier word of a governing clause. (See advcl: adverbial clause modifier, 2021a; advcl: adverbial clause modifier, 2021b)	advcl

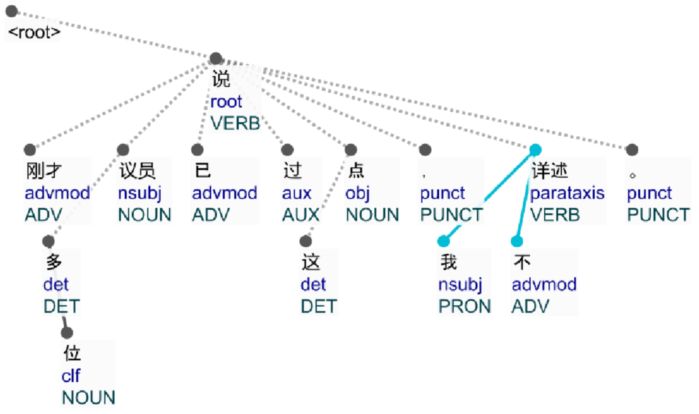
	 <root> 提出 root VERB ， punct PUNCT 可以 aux AUX 法庭 obl NOUN 质疑 obj NOUN 。 punct PUNCT 同意 advcl VERB 向 case ADP 如果 mark SCONJ 你 nsubj PRON 不 advmod ADV 决定 obj NOUN 我 nmod PRON 的 case PART 如果你不同意我的决定，可以向法庭提出质疑。 *Rúguǒ nǐ bù tóngyì wǒ de juédìng, kěyǐ xiàng fǎtíng tíchū zhìyí.* ‘If you disagree with my decision, you can challenge it in court.’ (sentence ID 952, HK-P treebank)	
Clausal modifier of noun	The clause represents adnominal dependent which modifies a noun. The clause can also occur in the form of depictives which are considered to be reduced non-verbal clauses. The clause can precede the noun (with or without 的, *de*, grammatical particle) or follow it without a function word between them. (See acl: clausal modifier of noun (adnominal clause), 2021a; acl: clausal modifier of noun, 2021b)  <root> 什么 root PRON 刚才 advmod ADV 问题 nsubj NOUN 是 cop AUX ？ punct PUNCT 留意 acl VERB 最大 amod ADJ 我 nsubj PRON 到 compound:vv VERB 的 mark:rel PART 刚才我留意到最大的问题是什么？ *Gāngcái wǒ liúyì dào zuìdà de wèntí shì shénme?* ‘What is the biggest issue which I just noticed?’ (sentence ID 846, HK-P treebank)	acl

All the examples of the UD trees were created by CoNLL-U Viewer (CoNLL-U Viewer) and adjusted by the author.

The clausal syntactic relation can also occur in a special form of parataxis (Table 4).

Table 4. Overview of a clausal extension – parataxis.

<table>
<tr><th>Clause</th><th>Description & Example</th><th>Tag</th></tr>
<tr><td rowspan="2">Parataxis</td><td>Parataxis is usually used for relation between a predicate and another clausal element which are placed next to each other without any further specification of their coordinate, subordinate or argument relation. The inventory of parataxis includes: side-by-side sentences separated by a colon or semicolon or placed next to each other without punctuation marks or a linking word, reported speeches without subordinate clausal structure, parenthetical comments, clausal interjections and tag questions.
(See parataxis: parataxis, 2021a; Parataxis: parataxis, 2021b)</td><td rowspan="2"><code>parataxis</code></td></tr>
<tr><td>

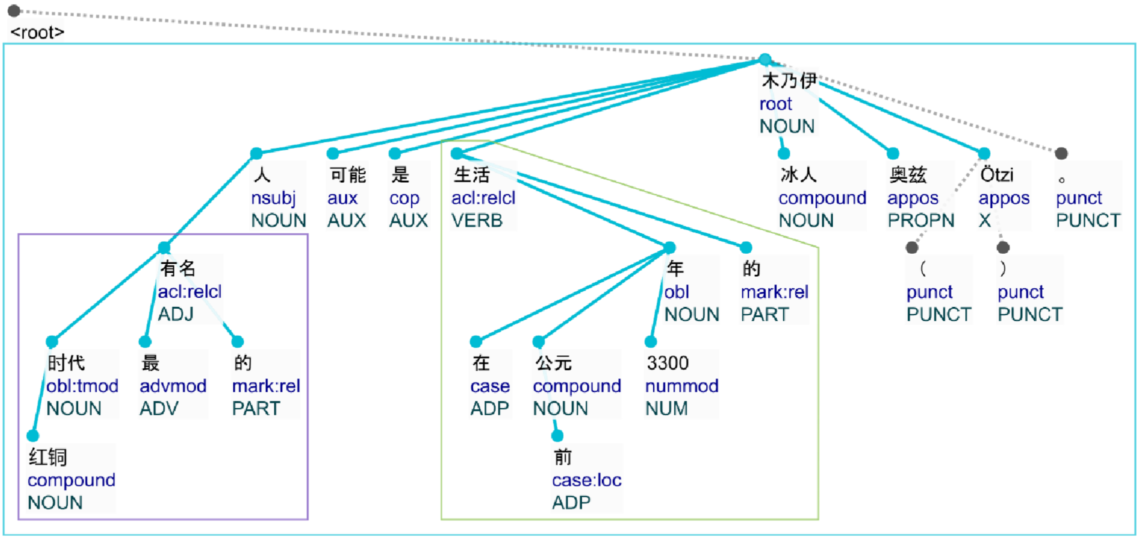


刚才多位议员已说过这点，我不详述。
Gāngcái duō wèi yìyuán yǐ shuō guò zhè diǎn, wǒ bù xiángshù.
‘Many members of the legislative body have already made this point, I will not go into details.’
(sentence ID 998, HK-P treebank)</td></tr>
</table>

The example of the UD tree was created by CoNLL-U Viewer (CoNLL-U Viewer) and adjusted by the author.

The subordinate clause can be operationalised in two different ways. The first approach regards the subordinate clause as an integral part of its governing clause and a separate clause at the same time. To illustrate the approach, we use a sentence from the PUD treebank as an example (Figure 4). When respecting the linear order of the sentence, the first subtree (framed in the violet box) identified by its clausal head 有名 (`acl:relcl`; *yǒumíng*, ‘well-known’) would be processed. The second subtree (framed in the green box) governed by the clausal head 生活 (`acl:relcl`; *shēnghuó*, ‘to live’) would follow. Finally, the processing would be carried out on the whole tree (framed in the blue box), which includes the main clause governed by the head 木乃伊 (`root`; *mùnǎiyī*, ‘mummy’) and the two previous clauses integrated into it as adnominal dependents modifying the subject (人, nsubj; *rén*, ‘person’) and the root (木乃伊,

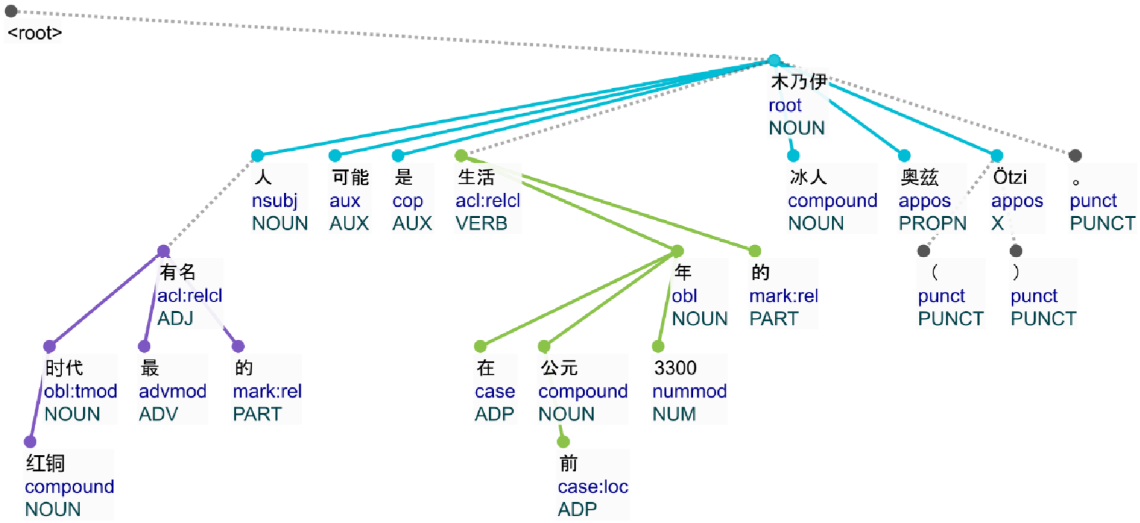
root; *mùnǎiyī*, ‘mummy’). This inclusive approach, however, results in multiple processing of the same sentential segments.



红铜时代最有名的人可能是生活在公元前3300年的木乃伊冰人奥兹（Ötzi）。
Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).
‘Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.’
Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author.

Figure 4. The example of the inclusive approach to clauses (sentence ID w02008038, PUD treebank). Tree nodes and edges belonging to the same clause are framed in a box of a given colour.

The second approach 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. Using the previous sentence as the example (Figure 5), the first two clauses (highlighted in violet and green) would be processed in the same manner, while the treatment of the last clause (highlighted in blue) would differ. Both its edges, i.e. between 1) 有名 (acl:relcl; *yǒumíng*, ‘well-known’) and 人 (nsubj; *rén*, ‘person’) and between 2) 生活 (acl:relcl; *shēnghuó*, ‘to live’) and 木乃伊 (root; *mùnǎiyī*, ‘mummy’), would be ignored (illustrated by dotted grey lines) and the clause would be treated only as a given subtree. 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. For this reason, we utilise only the second approach.



红铜时代最有名的人可能是生活在公元前3300年的木乃伊冰人奥兹（Ötzi）。
Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).
'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'
Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author.

Figure 5. The example of the exclusive approach to clauses (sentence ID w02008038, PUD treebank). Tree nodes and edges belonging to the same clause are highlighted in the same colour.

Let us compare our approach with other studies on Chinese. The determination of the clause followed by the thesis is similar to the determination used by Berdicevskis (2021). Both approaches differ only in the treatment of the conjunct (`conj`) and the open clausal complement (`xcomp`). Berdicevskis (2021) processed `conj` as a clausal dependency relation if a word with the `conj` tag or its governor was a verb, and he treated `xcomp` as a clausal dependency relation only if a word carrying this tag was a verb. However, an objection to his study can be raised – results of Chinese appear to be biased. A sample tested by Berdicevskis (2021) included both versions of the largest Chinese UD treebank (i.e. GSD), which differ only in the usage of various forms of Chinese characters (i.e. traditional and simplified).[100] Hence, each sentence was double processed while the form of the Chinese characters did not have any impact on the results of syntactic levels under analysis. The majority of other studies on Chinese determined the clause based on selected punctuation marks (Bohn, 1998, 2002; Hou et al., 2017; Jin and Liu, 2017; Chen, 2018; Chen and Liu, 2019, 2022; Hou et al., 2019a, 2019b; Sun and Shao, 2021). Some authors argued in favour of this approach because a segment between two punctuation marks approximates the clause (e.g. Jin and Liu, 2017; Chen and Liu, 2019, 2022). However, such a determination does not have to be grammatically exact (Chen, 2018; Chen and

[100] To compile a final sample for a given language, each treebank was conditioned to contain at least 10k tokens. As for Chinese, only the GSD and PUD treebanks should satisfy the condition. However, the number of sentences processed by Berdicevskis (2021) for Chinese exceeded 11k (AleksandrsBerdicevskis/menzerath/data_means/Chinese_sent.tsv, 2021) which implies inclusion of both the versions of GSD treebank (GSD contains ca 4k sentences and PUD 1k sentences).

Liu, 2019, 2022). It is noteworthy that Western-derived punctuation was integrated into Chinese relatively recently. Such efforts mainly appeared in the ’20s and ’30s of the 20th century (e.g. Mullaney, 2017). Hence, its usage might not be still stabilised and could even differ among authors (e.g. lead to its overuse, Hou et al., 2017).

3.2.3 The syntactic phrase

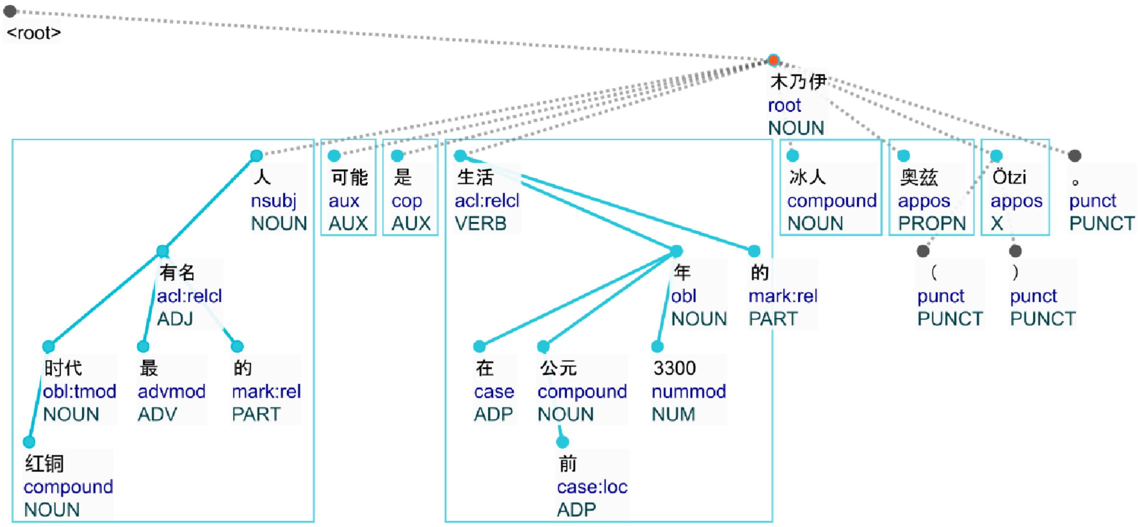
In general, the syntactic phrase (or shortly phrase) represents any subtree starting with a word (a node) being a phrasal head and continuing with other – directly or indirectly – dependent words (nodes). Regarding its determination, we firstly follow an approach introduced by Mačutek, Čech and Milička (2017). As mentioned in Chapter 2.1.1.2, 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.[101] Firstly, we precisely follow Mačutek, Čech and Milička (2017), i.e. only phrases directly depending on a head of a sentence (i.e. a root) are taken into account. Secondly, we apply the same approach to all clausal heads identified within a sentence (Berdicevskis, 2021).

In the case of the first approach, the syntactic phrase is viewed as a complete subtree – 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. As an illustration (Table 5 and Figure 6), a sentence from the PUD treebank is used and decomposed into seven phrases which directly depend on the root 木乃伊 (`root`; *mùnǎiyī*, ‘mummy’).

[101] 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 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, 2019a), hence, we do not go into the depth and take the coordination into account on the phrasal level.

Table 5. The example of sentential phrases in a sentence (sentence ID w02008038, PUD treebank).

ID	Phrase in characters	Phrase in pinyin	Translation into English
1	红铜时代最有名的人	*hóngtóng shídài zuì yǒumíng de rén*	'the most well-known person from the Copper Age'
2	可能	*kěnéng*	'probably'
3	是	*shì*	'is'
4	生活在公元前 3300 年的	*shēnghuó zài gōngyuán qián 3300 nián de*	'who lived during 3300 years B.C.'
5	冰人	*bīngrén*	'ice-man'
6	奥兹	*Àozī*	'Ötzi'
7	Ötzi	-	-



红铜时代最有名的人可能是生活在公元前 3300 年的木乃伊冰人奥兹（Ötzi）。

Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).

'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'

Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author

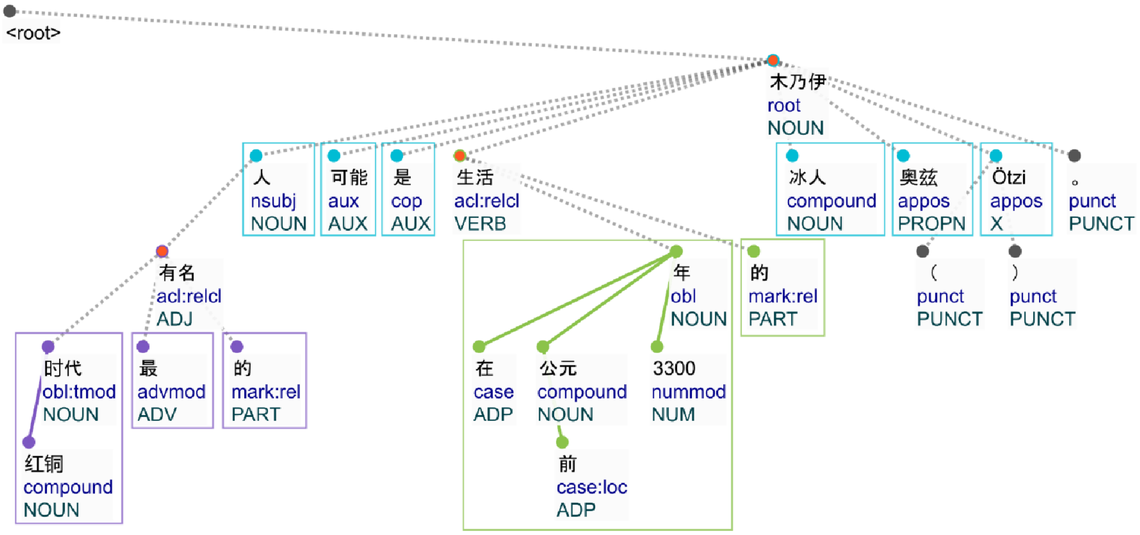
Figure 6. The example of sentential phrases in a sentence (sentence ID w02008038, PUD treebank). Each box frames one sentential phrase.

The second approach treats the phrase as a subtree that hangs from the head of each simple clause. The phrase cannot be the clause itself,[102] 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' and illustrate it with the same sentence used in the example above. Firstly, heads of simple clauses are identified (highlighted in orange in Figure 7): 1) 有名 (`acl:relcl`; *yǒumíng*, 'well-known'), 2) 生活 (`acl:relcl`; *shēnghuó*, 'to live'), 3) 木乃伊 (`root`; *mùnǎiyī*, 'mummy'). Secondly, subtrees directly dependent on the heads are determined and checked whether they are not clauses themselves (e.g. the subtree governed by the word 生活, *shēnghuó*, 'to live', is disregarded as the clausal phrase of the root 木乃伊, *mùnǎiyī*, 'mummy') or whether they do not contain another clause (e.g. the clausal phrase of the root 木乃伊, *mùnǎiyī*, 'mummy', governed by the word 人, *rén*, 'person', is reduced by a clause which it contains). As we can see (Table 6 and Figure 7), the inventory of the phrases changed.

Table 6. The example of clausal phrases in a sentence (sentence ID w02008038, PUD treebank).

ID	**Phrase in characters**	**Phrase in pinyin**	**Translation into English**
1	红铜时代	*hóngtóng shídài*	'copper Age'
2	最	*zuì*	'the most'
3	的	*de*	'grammatical particle'
4	人	*rén*	'person'
5	可能	*kěnéng*	'probably'
6	是	*shì*	'Is'
7	生活在公元前 3300 年	*shēnghuó zài gōngyuán qián 3300 nián*	'lived during 3300 years B.C.'
8	的	*de*	'grammatical particle'
9	冰人	*bīngrén*	'ice-man'
10	奥兹	*Àozī*	'Ötzi'
11	Ötzi	-	-

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



红铜时代最有名的人可能是生活在公元前 3300 年的木乃伊冰人奥兹（Ötzi）。

Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).

'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'

Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author

Figure 7. The example of clausal phrases in a sentence (sentence ID w02008038, PUD treebank). Phrases belonging to the same simple clause are framed in boxes of the same colour.

Both the approaches, however, have their drawbacks. In the case of the first approach, sentences whose roots do not govern any phrases have lengths equal to zero because the root is not the phrase itself (e.g. 木乃伊, root, *mùnǎiyī*, 'mummy', highlighted in orange in Figure 6).[103] Hence, these sentences are excluded from the analysis. In the case of the clausal phrase, similarly, clauses consisting only of its heads are not analysed when the clause becomes the construct (e.g. 有名, acl:relcl, *yǒumíng*, 'well-known' highlighted in orange in Figure 7). The question arises of how to treat these clauses of zero lengths with respect to sentences. One method might be to include clauses without phrases in the sum of all clauses in a sentence. However, the number of phrases would remain unchanged. As a result, the mean clause length would be lowered. Another method might be to disregard these clauses without phrases completely, i.e. they would not be included in the sum of the clauses in the sentence. The mean clause length would not be lowered in this case, but even more nodes would be left out of the analysis. The question is how much these methods influence the results when the law is applied. Hence, we test both. The clausal phrase faces another methodological difficulty. Due to the criterion that the phrase must not be the clause itself, not all dependency relations, i.e. edges, between the clausal head and its directly dependent elements are taken into account. For example, the relation between 木乃伊 (root, *mùnǎiyī*, 'mummy') and 生活 (acl:relcl,

[103] The sentential phrase can be treated with other alternative methods, e.g. by including the root in each sentential phrase (cf. Mačutek, Čech and Courtin, 2021). Nevertheless, these methods also have their drawbacks. Since we do not test them, we do not go into further details.

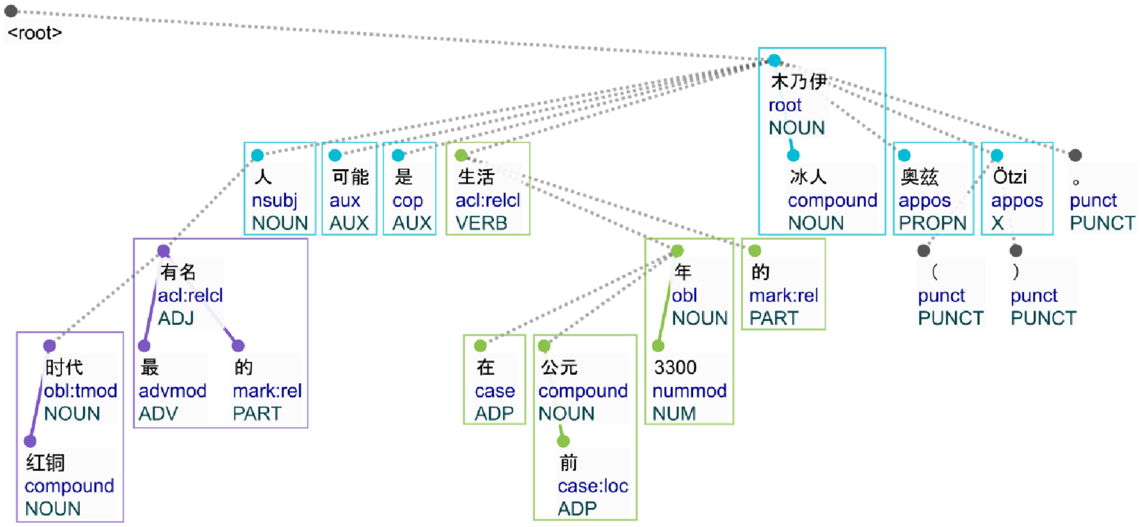
shēnghuó, ‘to live’) is neglected (depicted by the dotted grey line in Figure 7). If the inclusive approach is applied, all the elements directly dependent on the clausal head would be considered but at the cost of multiple processing of the same sentential segment.

As mentioned in Chapter 2.1.1.3, the approach by Mačutek, Čech and Milička (2017) was later revised by Mačutek, Čech and Courtin (2021). The authors discussed its drawbacks not only from the perspective of the predicate’s exclusion or inclusion but also in connection with a) phrasal lengths being above a threshold of the short-term memory and b) the linear property of language being ignored. For this reason, the authors suggested an alternative approach which determines 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).[104] The authors term the unit as a linear dependency segment (LDS) and we illustrate it with the same sentence used in the previous examples (see Table 7 and Figure 8). Based on results from Czech dependency treebanks, Mačutek, Čech and Courtin (2021) tentatively concluded that LDS might be a legitimate language unit, but it needs to be tested on other typologically different languages and on triplets of units where LDS occupies different positions than the one being analysed. Nevertheless, the approach appears to be overcoming the difficulties of the sentential and clausal phrase described above.

[104] 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.

Table 7. The example of linear dependency segments in a sentence (sentence ID w02008038, PUD treebank).

ID	LDS in characters	LDS in pinyin	Translation into English
1	红铜时代	*hóngtóng shídài*	'copper Age'
2	最有名的	*zuì yǒumíng de*	'the most well-known (+ grammatical particle)'
3	人	*rén*	'person'
4	可能	*kěnéng*	'probably'
5	是	*shì*	'is'
6	生活	*shēnghuó*	'lived'
7	在	*zài*	'during'
8	公元前	*gōngyuán qián*	'B.C.'
9	3300年	*3300 nián*	'3300 years'
10	的	*de*	'grammatical particle'
11	木乃伊冰人	*mùnǎiyī bīngrén*	'frozen mummy'
12	奥兹	*Àozī*	'Ötzi'
13	Ötzi	-	-



红铜时代最有名的人可能是生活在公元前3300年的木乃伊冰人奥兹（Ötzi）。

Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).

'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'

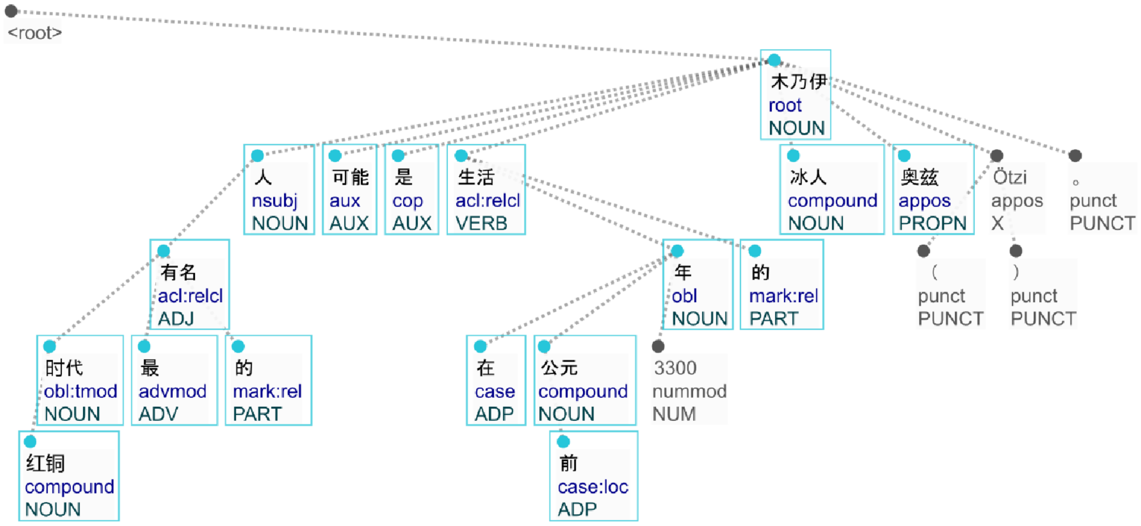
Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author

Figure 8. The example of linear dependency segments in a sentence (sentence ID w02008038, PUD treebank). LDSs belonging to the same simple clause are framed in boxes of the same colour.

The only study that tested the syntactic phrase in Chinese was published by Berdicevskis (2021), who followed the determination proposed by Mačutek, Čech and Milička (2017) and analysed the clausal phrases. Apart from the objection to his methodology raised above (see Chapter 3.2.2), we lack information on how the author dealt with issues related to potential multiple processing of the same sentential segments and clausal heads without phrases. Other studies on Chinese which did not intentionally integrate the phrase into menzerathian hierarchies of language units justified the exclusion by the difficult phrase determination in Chinese (Chen and Liu, p. 2, 2019, 2022, p. 4). Additionally, the studies argued that the phrase might be dispensable because of the neighbourhood between the clause and word (Chen and Liu, 2019, p. 7) or its approximation to the clause (Sun and Shao, 2021, p. 36). Due to the lack of ample evidence, the thesis tests all the phrases introduced above – sentential, clausal and LDS – to shed light on their behaviour when occupying different positions in the menzerathian hierarchy of language units in Chinese.

3.2.4 The word

UD and its annotation build on dependency relations between words (de Marneffe et al., 2021, p. 257; Tokenisation and Word Segmentation, 2021), representing nodes in a dependency tree and carrying morphosyntactic annotation (see Figure 9).



红铜时代最有名的人可能是生活在公元前 3300 年的木乃伊冰人奥兹（Ötzi）。
Hóngtóng shídài zuì yǒumíng de rén kěnéng shì shēnghuó zài gōngyuán qián 3300 nián de mùnǎiyī bīngrén Àozī (Ötzi).
'Likely the most well-known person from the Copper Age is Ötzi, the frozen mummy who lived during 3300 BC.'
Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author.

Figure 9. The example of words in a sentence (sentence ID w02008038, PUD treebank). Tree nodes considered to be the words under analysis are framed in blue boxes.

The word segmentation in the UD is driven by algorithms which are specific to a given language. In the case of the Chinese-HK UD treebank (Poiret et al., 2021, pp. 5-6), the algorithm follows segmentation guidelines (Xia, 2000) which was developed for the Chinese Treebank[105] (Xue et al., 2013), a large Chinese corpus using phrase structure annotation[106]. Roughly speaking, the guidelines see the word as a basic syntactic element, called a syntactic atom (Xia, 2000, p. 5). Due to certain factors which complicate the determination of word boundaries in Chinese (i.e. a lack of spaces between words, minimal inflection or disagreement on segmentation of complicated constructions), the guidelines utilise several rules to identify the word, i.e. 1) a bound morpheme is a part of a word, 2) segmentation of complex internal structures is preferred, 3) the meaning of morphemes in a word is not compositional, 4) morphemes of a word cannot

[105] Formerly known as the Penn Chinese Treebank.

[106] The dependency (DS) and phrase (PS) structures differ mainly in the form of relations (being between a parent and child in DS and between siblings in PS), syntactic structures (which is verb central in DS and binarily divided into two phrases in PS), correspondence between words and nodes (being one-to-one in DS while PS also allows one-to-many) and headedness (each node has only one governor except for a root in DS which is not necessary for PS), (Osborne, 2019b, pp. 362-365).

be separated by insertion of another morpheme, 5) a morpheme of a word is not replaceable by a phrase, 6) segmentation can be driven by the number of syllables (Xia, 2000, pp. 4-5). The Chinese-HK UD treebank only diverges from the Chinese Treebank in the treatment of verbal compound structures – resultative compounds (verb + resultative suffix, e.g. 做好, *zuòhǎo*, 'done' or 'finished') and potential complement (verb + potential complement + verb, e.g. 买不到, *mǎibudào*, 'be unable to buy'), (Poiret et al., 2021, p. 6). The former always splits them into separate word tokens while the latter treats them variously based on the rules. Information on the word segmentation in GSD and PUD is not available. In the case of LCMC, the word segmentation was performed using the Chinese Lexical Analysis System ICTCLAS (Institute of Computing Technology of Chinese Academy of Science, n.d.).

Generally, the Chinese word in the UD and LCMC samples corresponds to a string of Chinese – traditional or simplified – characters. However, the samples also include words which partly or entirely consist of non-Chinese graphemes, i.e. letters of the Latin alphabet, Arabic numerals and/or other symbols (e.g. 3000 or Ötzi in our example, highlighted in grey in Figure 9).[107] We treat these words differently with respect to their position in the unit hierarchy. If such a word is the direct constituent to higher linguistic levels in the UD treebanks, we include it in the analysis because it occupies a syntactic position which cannot be left out. On the other hand, one Chinese grapheme, i.e. character, roughly corresponds to a syllable, whereas one non-Chinese grapheme mostly represents a letter or numeral. Hence, both the types do not correspond to each other, which amplifies the heterogeneity of the samples. For this reason, we also test the exclusion of those constructs in which words partly or entirely consist of non-Chinese graphemes. If such a word is the construct itself, we directly exclude it from the analysis of the UD treebanks and LCMC due to a higher degree of homogeneity. Punctuation marks are also annotated in the UD treebanks and LCMC (e.g. a full stop '。' and parentheses '（）' highlighted in grey in Figure 9). However, we do not take them into account at any level (even though there are studies which treat them as language units, e.g. Hug, 2004, Benešová and Birjukov, 2015).

Finally, the frequency of the word being the construct must be considered. As discussed in Chapter 1.4, tokens and types have different impacts on results. The tokens reflect the frequency of use and seem to be governed by the Brevity law rather than the Menzerath-Altmann law. The usage of shorter units is preferred in this case which might prevent the Menzerath-Altmann law from coming into force. On the contrary, the types should not be biased in this way. When taking a look at the results on the word level in Chinese, the law was corroborated only by Bohn (1998, 2002), who tested word types. Results of the word tokens were in contradiction to the law (Motalová and Matoušková, 2014; Chen and Liu, 2016, 2019, 2022) unless a direct measurement unit of the word (i.e. Chinese character) was skipped (Chen and Liu, 2016, 2019, 2022). In addition, the tokens and the types have not been tested on the same language material in Chinese. The thesis fills the gap and the law is applied to both.

[107] The non-Chinese words or words that mix both – Chinese and non-Chinese – graphemes account for 0.19 % of word tokens and 0.89 % of word types in HK-P, 3.60 % of word tokens and 8.86 % of word types in PUD, and 3.03 % of word tokens and 9.24 % of word types in GSD.

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 of 6,647 Chinese characters.[108] However, to use the document, all words in the samples must be written in simplified Chinese characters. LCMC consists of texts written in simplified characters and the UD Chinese GSDSimp treebank is already a result of automatic conversion and manual correction performed by the UD project itself. Only the UD Chinese HK and PUD treebanks contain words written in traditional Chinese characters. For this reason, we had to convert these two treebanks into their simplified forms by virtue of available software (文林 Wénlín Software for Learning Chinese: Version 4.0.2, 2011).

The simplified Chinese characters and their strokes are immutable. Hence, their determination is the same across studies. On the other hand, approaches to the component lack a consensus. Bohn (1998, 2002) and Chen and Liu (2016, 2019, 2022) used different sources containing the decomposition of Chinese characters into their components, while Motalová et al. (2013), Motalová and Matoušková (2014), Matoušková and Motalová (2015) and Matoušková (2016) introduced their determination.

As for the Chinese character being the word direct constituent, Chen and Liu (2019, 2022) opted for the component and stroke to be both its measurement units. The authors did not corroborate the law for any of the triplets. However, they tested only word tokens. For this reason, we decided to follow their approach and test the influence of both the units on the Chinese character when not only the word tokens but also the word types are analysed.

Lastly, similarly to the word, we consider the frequency factor and analyse the tokens and the types with respect to the Chinese characters. The analysis of the tokens prevails while the types were analysed only by Bohn (1998, 2002) and no study, to our best knowledge, tested both.

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,

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

we disregard quantitative differences between characters and syllables because erization occurs in our samples to a minimal extent.[109]

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 (see Table 8), 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. The applied rules are as follows:

1. Firstly, the post-alveolar affricative <ch, zh>, fricative <sh> and the velar nasal <ng> consonants are captured by two letters in pinyin, while in IPA being only one sound [tʂʰ], [tʂ], [ʂ] and [ŋ] respectively. Hence, the digraphs are reduced to one symbol.
2. Secondly, "labial consonant cannot be followed by a mid vowel in CV [consonant-vowel] syllable" (Lin, 2007, 119). Therefore, if the syllable starts with <b>, <p>, <m> and <f>, the vowel <o> in pinyin is prolonged by one symbol to correspond two sounds in IPA [wo], i.e. <bo>, <po>, <mo>, <fo> vs [bwo], [pʰwo], [mwo], [fwo] (Lin, 2007, p. 128).
3. Thirdly, the rules have an impact on the diphthongs <ai>, <ao>, <ei> and <ou>, each of which is viewed as a complex vowel modifying only its quality in a syllable (Lin, 2007, p. 69), or in other words, as one sound. For this reason, the digraphs are reduced to one symbol.
4. Fourthly, another quantitative difference is caused by the schwa (Lin, 2007, p. 127). The algorithm inserts it in syllables where <i> / [j] is preceded by a consonant and directly followed by the velar nasal <ng> / [ŋ] (e.g. <bing> vs [bjəŋ]).[110] The insertion of the schwa is also applied to those syllables where <u> / [w] is preceded by a consonant different from the alveolo-palatals <j> / [tɕ], <q> / [tɕʰ] and <x> / [ɕ], and followed by the alveolar nasal <n> / [n] (e.g. <dun> vs. [twən]).[111]
5. Lastly, if <yu> / [ɥ] occupies the initial position of a syllable and precedes <e> / [e] or <an> / [ɛn], the syllable length changes from <yue> and <yuan> to [ɥe] and [ɥɛn] accordingly (Lin, 2007, p. 129).

[109] 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.

[110] The only syllable not affected by the rule is <ying> / [jəŋ] because 1) it starts with a semi-vowel (called glide) and 2) there is no quantitative difference in length of <yi> and [jə].

[111] The schwa is not inserted if a syllable starts with the semi-vowel (glide) <y> / [ɥ].

Table 8. 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

Due to the close correspondence between sounds and pinyin letters, we do not expect considerable differences between the syllable lengths measured in sounds and pinyin letters. Hence, we determine the length of the syllable only as a sequence of a letter(s) and/or symbol(s) used for the pinyin alteration (Table 9), or in other words, as a sequence of sounds.

Table 9. The example of pinyin and its alternation corresponding to sounds in IPA (sentence ID w02008038, PUD treebank).

Word	Number of syllables	Syllables in pinyin	Number of letters	Syllables in sounds	Number of sounds	Difference*
红铜	2	['hong', 'tong']	8	['hoŋ', 'toŋ']	6	-2
时代	2	['shi', 'dai']	6	['$i', 'd#']	4	-2
最	1	['zui']	3	['zui']	3	0
有名	2	['you', 'ming']	7	['y#', 'mjəŋ']	6	-1
的	1	['de']	2	['de']	2	0
人	1	['ren']	3	['ren']	3	0
可能	2	['ke', 'neng']	6	['ke', 'neŋ']	5	-1
是	1	['shi']	3	['$i']	2	-1
生活	2	['sheng', 'huo']	8	['$eŋ', 'huo']	6	-2
在	1	['zai']	3	['z#']	2	-1
公元	2	['gong', 'yuan']	8	['goŋ', 'ɥæn']	6	-2
前	1	['qian']	4	['qian']	4	0
年	1	['nian']	4	['nian']	4	0
的	1	['de']	2	['de']	2	0
木乃伊	3	['mu', 'nai', 'yi']	7	['mu', 'n#', 'yi']	6	-1
冰人	2	['bing', 'ren']	7	['bjəŋ', 'ren']	7	0
奥兹	2	['ao', 'zi']	4	['#', 'zi']	3	-1

*when IPA is compared to pinyin

The inventory of the Chinese syllables does not differ across studies on Chinese. However, differences occur when it comes to a phoneme or sound. Schusterová et al. (2013) and Ščigulinská and Schusterová (2014) determined the phoneme based on a Czech transcription, while Chen and Liu (2016) used a pronunciation list of the Chinese characters without any reference.

3.3 Language unit combinations and their quantification

As discussed in Chapter 1.4, it is assumed that the menzerathian tendency between the lengths of the construct and the constituent appears as far as immediately neighbouring units are concerned (e.g. Altmann, 1983; Cramer, 2005a). Hence, the choice of the direct constituent to the construct exerts a strong influence on results, although their neighbourhood is not always unambiguous. The following section introduces the measurement units, i.e. the 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 which correspond to syllables in Chinese only except for erization (as addressed in the previous Chapter 3.2.6).

Character

Measurement unit: component or stroke – The length of the simplified Chinese character is calculated either as a sum of its components, 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.

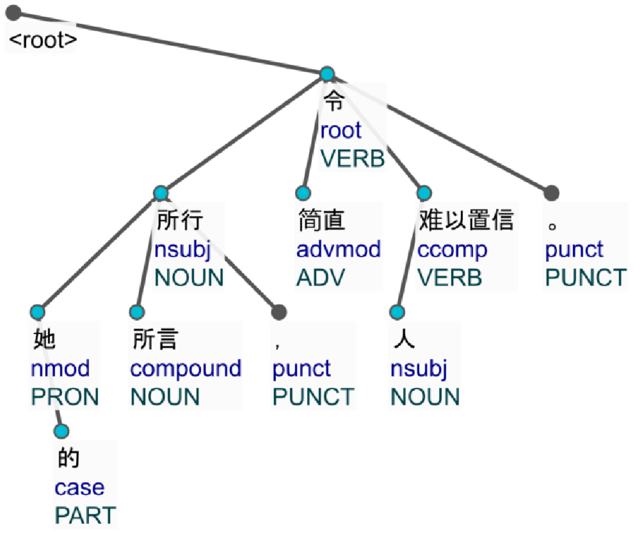
To analyse the menzerathian relationship, we always need a triplet of language units – construct, constituent and sub-constituent. The length of a unit of the highest position, i.e. a construct, is measured as a sum of lower units, i.e. constituents, from which the construct is directly constructed. The length of the constituent is measured in the lowest units in this hierarchy of three, or in other words, in its direct constituents or indirect constituents of the construct (i.e. sub-constituents). All analysed triplets are included in Table 10 with studies on Chinese which tested them. As can be seen, the law has been applied to the phrase and the word types to the least extent. In addition, the phrase and the linear dependency segment proposed by Mačutek, Čech and Milička (2017) and Mačutek, Čech and Courtin (2021) accordingly have been tested by the authors only with respect to a particular position (i.e. being constituent and sub-constituent of the sentence respectively), even though the unit can be integrated up to three triplets (i.e. being construct, constituent and sub-constituent). For this reason, the thesis tests the language units in all possible positions to shed light on their behaviour when the position is changed and the Chinese language is tested.

Table 10. 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)

*data not reliable

Finally, we exemplify the calculation of the construct and constituent lengths for each triplet in Table 11, Table 12 and Table 13 while using the same sentence from PUD (Figure 10) as an example.



她的所言所行，简直令人难以置信。

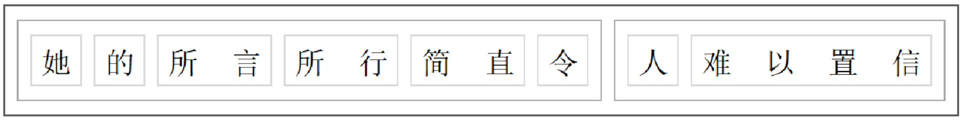
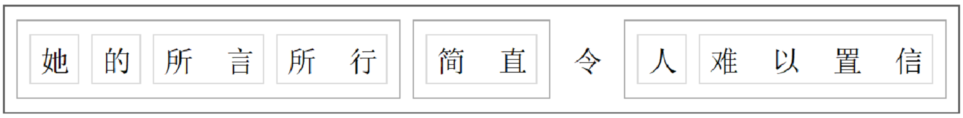
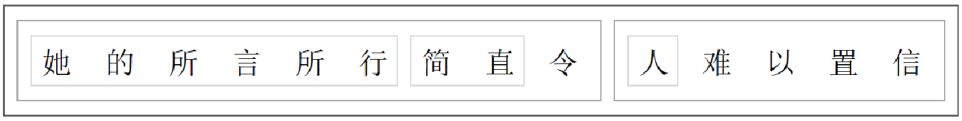
Tā de suǒyán suǒxíng, jiǎnzhí lìng rén nányǐzhìxìn.

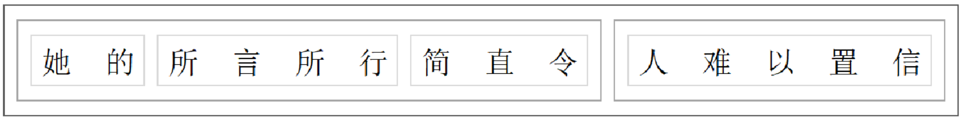
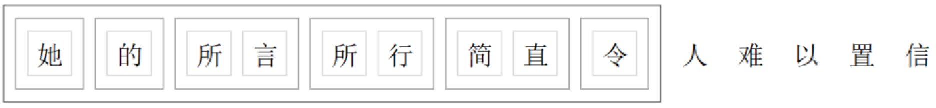
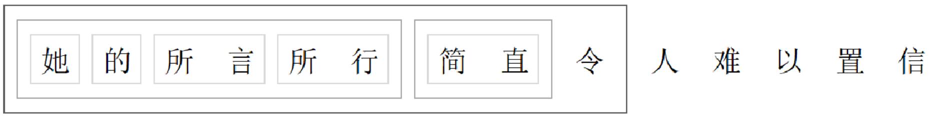
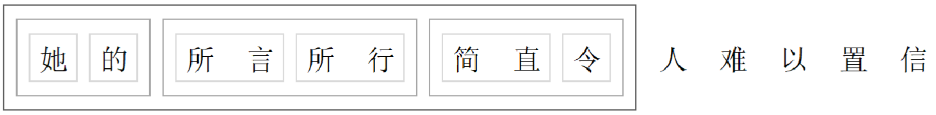
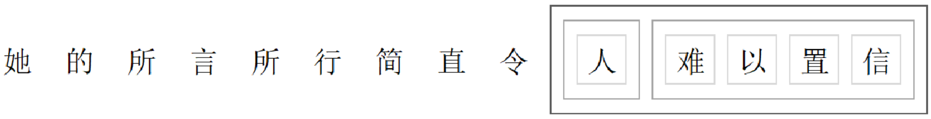
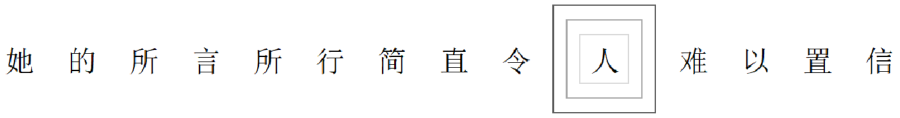
'What she is saying and what she is doing, it — actually, it is unbelievable.'

Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author.

Figure 10. The example of a sentence (sentence ID n01002058, PUD treebank).

Table 11. Calculation of unit lengths belonging to triplets on the syntactic level.

Construct x	**Constituent y**	**Sub-constituent**	**Length x and y**
Sentence	Clause		$x = 2$
		Word	$y = (6 + 2)/2 = 4.00$
			
Sentence	Sentential phrase		$x = 3$
		Word	$y = (4 + 1 + 2)/3 = 2.33$
			
Sentence	Clause		$x = 2$
		Clausal phrase	$y = (2 + 1)/2 = 1.50$
			

Sentence	Clause		$x = 2$
		LDS	$y = (3 + 1)/2 = 2.00$
			
Clause	Word		$x = 6$
		Character	$y = (1 + 1 + 2 + 2 + 2 + 1)/6 = 1.50$
			
Clause	Clausal phrase		$x = 2$
		Word	$y = (4 + 1)/2 = 2.50$
			
Clause	LDS		$x = 3$
		Word	$y = (2 + 2 + 2)/3 = 2.00$
			
Sentential phrase	Word		$x = 2$
		Character	$y = (1 + 4)/2 = 2.50$
			
Clausal phrase	Word		$x = 1$
		Character	$y = 1/1 = 1.00$
			

LDS	Word		$x = 2$
		Character	$y = (1 + 4)/2 = 2.50$
她 的 所 言 所 行 简 直 令 [人] [难 以 置 信]			

Table 12. Calculation of unit lengths belonging to triplets on the word level. The word 难以置信 (*nányǐzhìxìn*, 'be difficult to believe') is used.

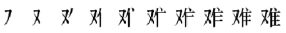
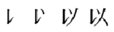
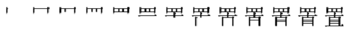
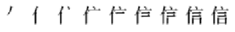
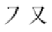
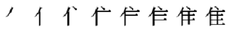
Construct x	**Constituent y**	**Sub-constituent**	**Length x and y**
Word	Character		$x = 4$
		Component	$y = (2 + 2 + 3 + 2)/4 = 2.25$
[又 隹] [厶 人] [罒 十 目] [亻 言]			
Word	Character		$x = 4$
		Stroke	$y = (10 + 4 + 13 + 9)/4 = 9.00$
Character 1: 难→  Character 2: 以→ 		Character 3: 置→  Character 4: 信→ 	
Word	Syllable		$x = 4$
		Sound	$y = (3 + 2 + 2 + 3)/4 = 2.50$
[n a n] [y i] [$ i] [x i n]			

Table 13. Calculation of unit lengths belonging to the triplet on the character level. The character 难 (*nán*, 'difficult') is used.

Construct x	**Constituent y**	**Sub-constituent**	**Length x and y**
Character	Component		$x = 2$
		Stroke	$y = (2 + 8)/2 = 5.00$
Component 1: 又→ 			
Component 2: 隹→ 			

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 for each triplet mentioned above. As addressed in Chapter 1.3, the low frequency of the constructs (mostly of the longest lengths) might result in irregular behaviour of its constituents. 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 (as applied on the syntactic level, e.g. by Köhler, 1982; Bohn, 1998, 2002; Mačutek, Čech and Milička, 2017; or on the word level, e.g. by Mačutek, Chromý and Koščová, 2018; Rujević et al., 2021), 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 (see Table 14).

Table 14. The example of the calculation of the weighted average. Original values of the construct length (x), its frequency (f), and the constituent length (y) are presented on the left, while the weighted values are on the right. Values to be pooled are highlighted in grey.

x	f	y
1	1989	2.39
2	2155	2.58
3	150	2.76
4	9	2.44
6	1	2.17

$$x_w = \frac{\sum_i^n x_i f_i}{\sum_i^n f_i} = \frac{(4 \times 9) + (6 \times 1)}{9 + 1} = 4.20$$

$$y_w = \frac{\sum_i^n y_i f_i}{\sum_i^n f_i} = \frac{(2.44 \times 9) + (2.17 \times 1)}{9 + 1} = 2.41$$

where x_w is the weighted average construct length, y_w the weighted average constituent length, x_i, y_i, f_i are values to be pooled.

x	f	y
1	1989	2.39
2	2155	2.58
3	150	2.76
4.2	10	2.42
-	-	-

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 choice is motivated by the possibility of the former model reflecting the second (or reverse) regime of the law while the latter model includes only one parameter, which eases its interpretability, as discussed in Chapter 1.3, 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 all the triplets introduced above. We divide the chapter first according to constructs and then according to its (direct and indirect) constituents. Each unit combination introduces a hypothesis which is followed by obtained results from all samples (HK-P, PUD, PUD-N, PUD-W, GSD and on the word and character level also from LCMC) presented in tables and figures. The figure includes graphs visualising the behaviour of obtained lengths and the fit of both the models for each sample separately. We use the same scale of both axes for all the samples to display their differences (if any). Next, we address and interpret the results and, if possible, apply an alternative approach. The sub-chapter on the construct ends with an overall summary.

As for the discussion, firstly, we comment on whether the constituents show an increase in their lengths, or in other words, the second (or reverse) regime. It should be emphasised that we consider the second regime in the strict sense, i.e. if any constituent length increases in comparison to its predecessor, even though we are fully aware that these increases might be only fluctuations from the overall decreasing trend.

Secondly, we comment only on the parameters of the truncated model, i.e. a and b, because the complete model lacks a linguistic interpretation of the parameter c. We compare values of each parameter across the samples and evaluate the relationship between both the parameters a and b if the coefficient of determination R^2 meets the standard of $R^2 \geq 0.90$ in most samples. Otherwise, we only address considerable changes in their values if different approaches are applied.

Thirdly, we follow studies (e.g. Jin and Liu, 2017; Jiang and Ma, 2020; Mačutek, Čech and Courtin, 2021) that assessed constituent lengths with respect to the short-term memory limit proposed by Miller (1956), i.e. 7 ± 2.[112] However, in Miller's view, it is not up to 7 ± 2 items that limit the short-term memory but rather 7±2 chunks resulting from a "process of organising or grouping the input into familiar units or chunks" (Miller, 1956, p. 93). Miller (1956, p. 93) illustrated the chunks with an example of a radiotelegraphic code – sounds as first-level chunks are grouped into letters, letters as second-level chunks into words and words as third-level chunks into phrases etc. This structure resembles the structure of constructs and constituents, or in other words, the menzerathian hierarchy of language units. Hence, we use Miller's limit to evaluate not only the constituent but also the construct.[113] The constituent length might meet the limit of the short-term memory span, but its construct might not.

Finally, we created scripts for data processing that are available on Github, where all processed data (including their non-weighted versions) can be found as well.[114]

[112] The concept of short-term (immediate or working) memory limits has been heavily discussed later (cf. Cowan 2000) while suggesting even a lower span, i.e. about four items (Cowan, 2000).

[113] In the case of the construct and constituent being pooled due to insufficient frequency, we evaluate their pooled value presented in this work.

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

4.1 The sentence as the construct

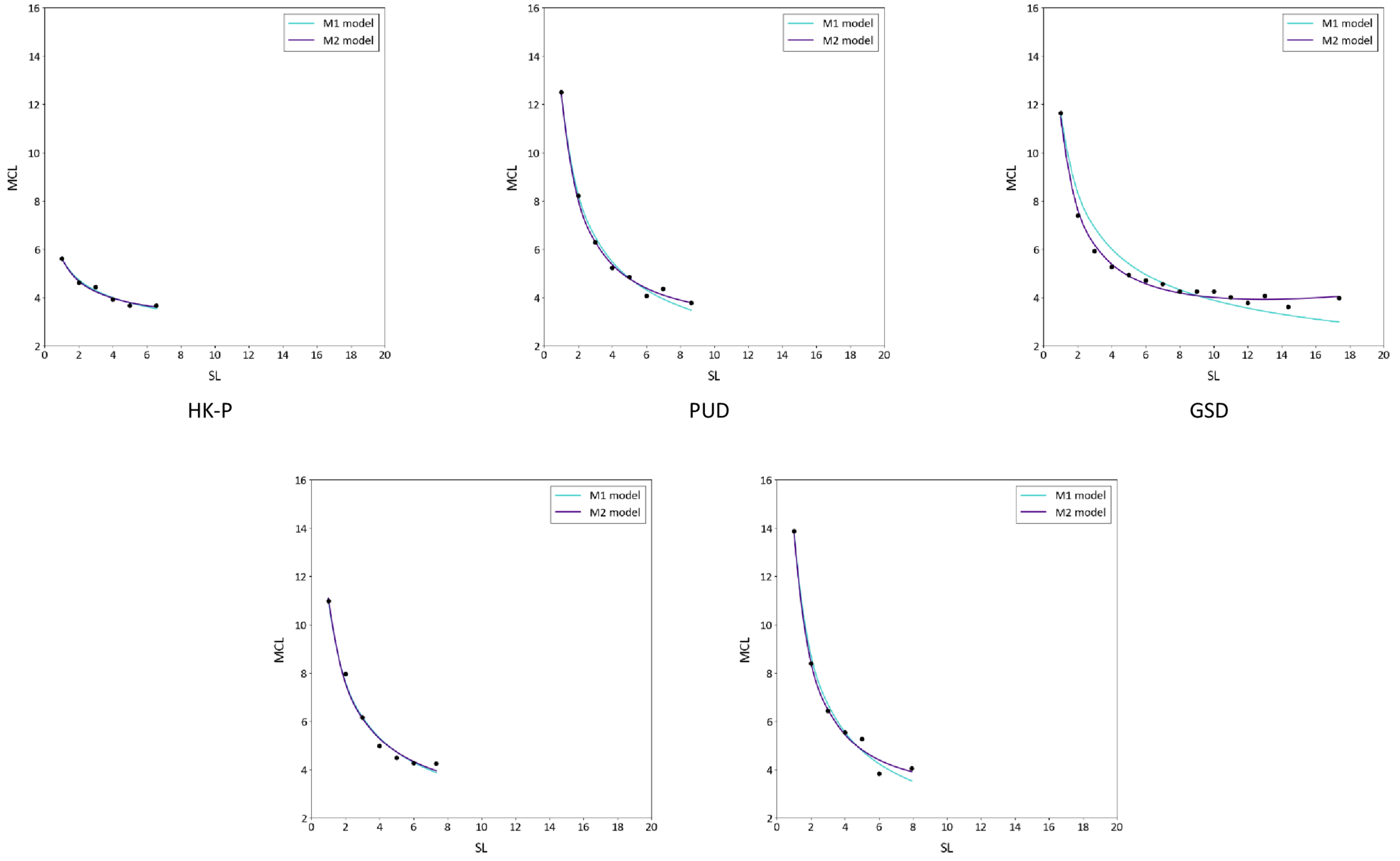
4.1.1 The clause and word as constituents

Hypothesis: the longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in words.

The results obtained by applying the law to all the samples are presented in Table 15 and Figure 11. SL denotes the sentence length measured the in the number of clauses, $f(SL)$ its frequency and MCL the mean clause length measured in the number of words. The table contains the parameters (a, b, c) and the coefficient of determination $R^2$of both models – the truncated model $y(x) = ax^b$ labelled as M1 and the complete model $y(x) = ax^b e^{cx}$ labelled as M2. In case of the former, the parameter a equals a value of the mean clause length of one-clause sentences, i.e. MCL_1. If a value of MCL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 15. MAL applied to the triplet of the sentence, clause and word.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*
1	75	5.63	1	175	12.50	1	83	10.99	1	92	13.87	1	407	11.64
2	97	4.62	2	271	8.21	2	115	7.97	2	156	8.39	2	840	7.39
3	73	4.43	3	248	6.29	3	133	6.17	3	115	6.44	3	830	5.93
4	56	3.92	4	140	5.24	4	77	4.98	4	63	5.55	4	636	5.28
5	23	3.67	5	84	4.85	5	46	4.50	5	38	5.27	5	446	4.95
6.57	30	3.67	6	51	4.08	6	28	4.28	6	23	3.83	6	290	4.70
			7	20	4.37	7.33	18	4.25	7.92	13	4.05	7	186	4.57
			8.64	11	3.79							8	137	4.26
												9	93	4.25
												10	39	4.25
												11	40	4.02
												12	18	3.79
												13	11	4.07
												14.38	13	3.61
												17.36	11	3.99
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	5.63	5.53	*a*	12.50	12.05	*a*	10.99	10.92	*a*	13.87	13.14	*a*	11.64	10.91
b	-0.25	-0.29	*b*	-0.59	-0.70	*b*	-0.52	-0.57	*b*	-0.66	-0.79	*b*	-0.48	-0.65
c		-0.02	*c*		-0.04	*c*		-0.02	*c*		-0.05	*c*		-0.05
R^2	0.9744	0.9779	R^2	0.9924	0.9964	R^2	0.9890	0.9897	R^2	0.9878	0.9923	R^2	0.9228	0.9915



HK-P PUD GSD

PUD-N PUD-W

Figure 11. MAL applied to the triplet of the sentence, clause and word.

The goodness-of-fit between both the models and the data reaches the standard of $R^2 \geq 0.90$. Hence, the hypothesis is not rejected for the triplet of the sentence, clause and word.

The second (or reverse) regime, i.e. mean clause lengths which increase in comparison to its predecessors (highlighted in yellow in Table 15), is observed only in curve tails of three out of five samples (i.e. in PUD, PUD-W and GSD). As Tanaka-Ishii pointed out, the "problem is that, for every point, the variation is usually very large. As a result, only the mean value exhibits a tendency to drop" (2021, p. 2). However, even the mean values do not have to decrease (as discussed in Chapter 1.3). The low frequency of a given construct length can lead to the deviation of its constituent from the menzerathian decreasing trend, and the law might not manifest itself compared to constituents of highly frequent constructs. Moreover, the less frequent construct lengths might possess specific properties (e.g. structure or content) which counteract the law.

The parameter a of M1 has the lowest value in HK-P ($a = 5.63$). In the case of the other samples, it reaches higher values and, in PUD-W, is the highest ($a = 13.87$). Its value appears to be under the influence of a) a linguistic level, i.e. measuring clauses in words leads to a higher variance in their lengths, and b) a text type which consequently comes into play. While PUD and GSD are of descriptive and informative nature (involving news and/or Wikipedia articles), HK-P inclines towards spoken nature (represented by proceedings), which usually shortens clausal lengths in words (as pointed out by several authors in connection with literary text types containing dialogues, e.g. Kułacka, 2009b, p. 27; Jin and Liu, 2017, p. 217; Hou et al., 2017, pp. 10-11). As for the parameter b of M1, the highest value is reached in HK-P ($b = -0.25$), where the shortening tendency of the fitting curve is minimal compared to the other samples, while the parameter in PUD-W reaches the lowest value ($b = -0.66$) and makes the slope of the curve the steepest. The values of both the parameters support the assumption of their negative correlation (Figure 12) – the higher the value of the parameter a, the lower the value of the parameter b (as confirmed e.g. by Hammerl and Sambor, 1993; Hou et al., 2019a; Jiang and Jiang, 2022).

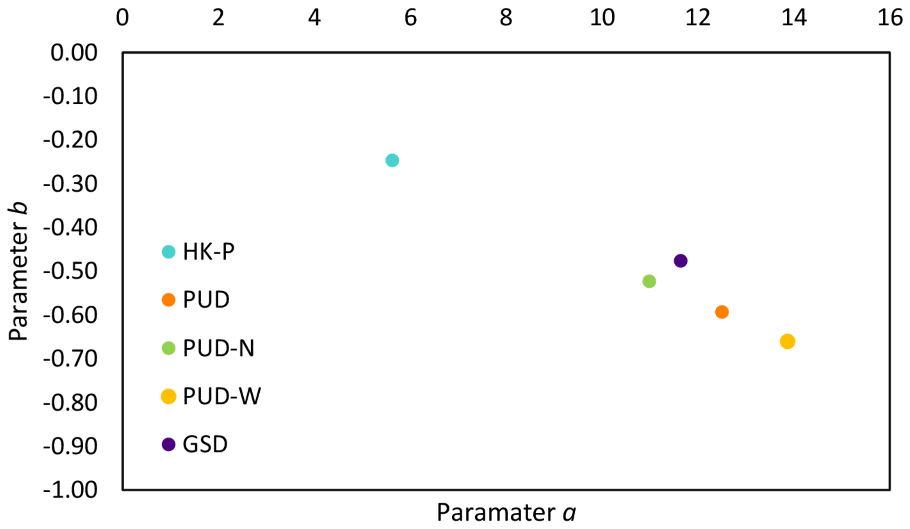


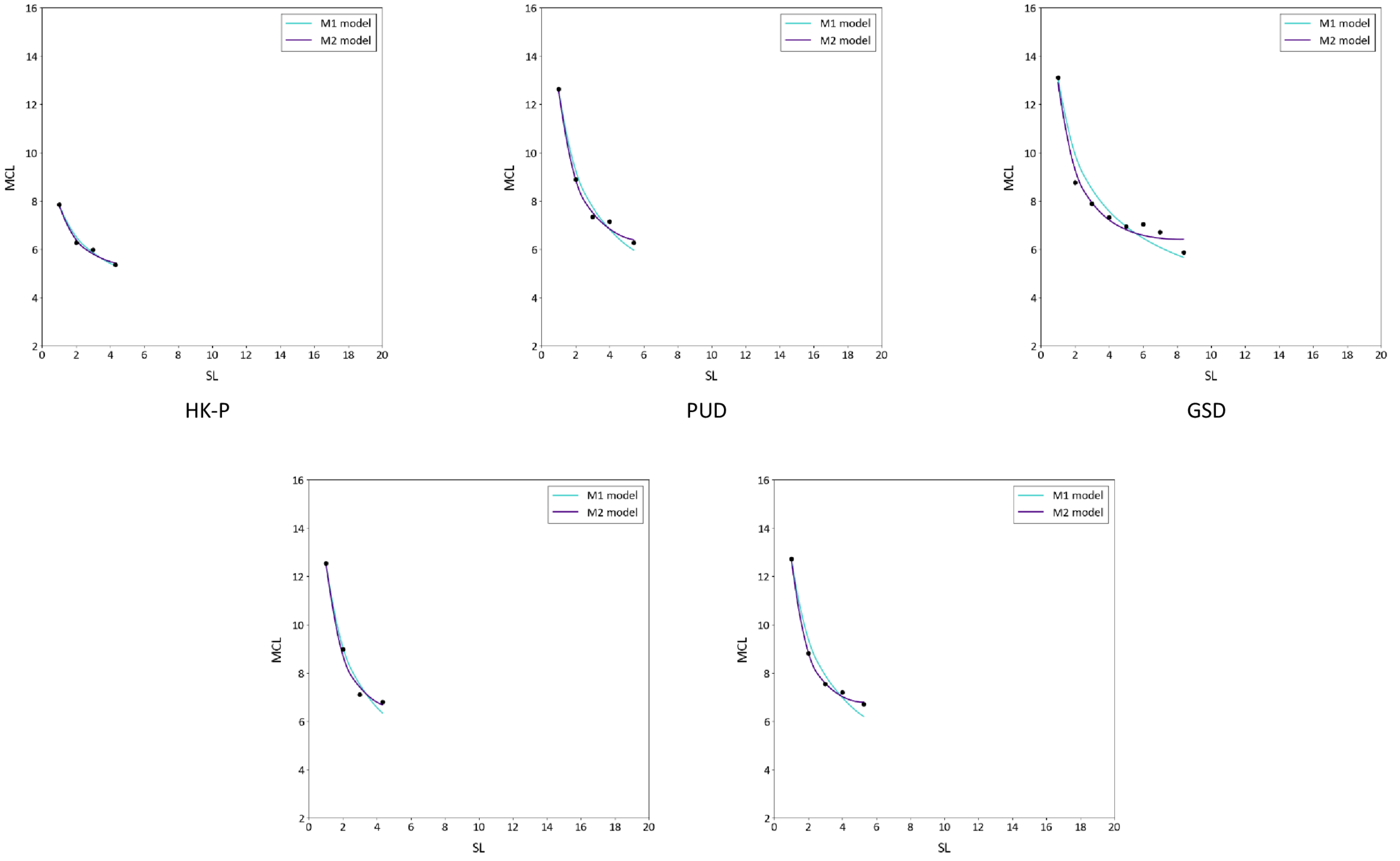
Figure 12. The parameters a and b of M1 for the triplet of the sentence, clause and word.

As for the construct, the scales of the sentence lengths obtained from HK-P, PUD and its versions do not exceed the upper short-term memory limit expressed by Miller’s number 7 ± 2 (Miller, 1956), while the scale from GSD does (i.e. $1 \leq SL \leq 17.36$). GSD differs from the other samples in size, which leads us to an assumption of higher variance in its sentence lengths.[115] Another factor to consider is the UD annotation for the clausal dependency relations. Hence, the question arises of whether an alternative approach to the clause would result in a different scale of sentence lengths. For comparison, we opt for the clause determination which was adopted by studies on Chinese and which relies on selected punctuation marks. We choose 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), and also extend the whole selection by an ellipsis ‘…’/’……’ (Sun and Shao, 2021). The obtained results are given in Table 16 and Figure 13.

[115] Almost 4k sentences in GSD versus 1k sentences in PUD and 354 sentences in HK-P.

Table 16. MAL applied to the triplet of the sentence, clause and word – punctuation approach.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*
1	140	7.85	1	249	12.63	1	134	12.55	1	115	12.73	1	756	13.10
2	138	6.27	2	435	8.90	2	219	8.98	2	216	8.81	2	1513	8.77
3	54	5.98	3	229	7.34	3	108	7.11	3	121	7.54	3	964	7.89
4.32	22	5.37	4	68	7.14	4.33	39	6.79	4	37	7.20	4	434	7.33
			5.42	19	6.28				5.27	11	6.71	5	180	6.95
												6	80	7.04
												7	41	6.71
												8.38	29	5.87
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	7.85	7.45	*a*	12.63	11.58	*a*	12.55	11.28	*a*	12.73	11.21	*a*	13.10	12.05
b	-0.27	-0.36	*b*	-0.44	-0.63	*b*	-0.47	-0.68	*b*	-0.43	-0.70	*b*	-0.39	-0.57
c		-0.05	*c*		-0.09	*c*		-0.11	*c*		-0.13	*c*		-0.07
R^2	0.9752	0.9850	R^2	0.9793	0.9946	R^2	0.9816	0.9927	R^2	0.9653	0.9982	R^2	0.9257	0.9734



HK-P PUD GSD

PUD-N PUD-W

Figure 13. MAL applied to the triplet of the sentence, clause and word – punctuation approach.

Firstly, the coefficient of determination R^2 of M1 and M2 achieved by the punctuation approach meets the standard of $R^2 \geq 0.90$ and the hypothesis is not rejected. Secondly, the punctuation approach mostly yields higher values of the M1 parameters a and b, which appear to be influenced by the determination of a linguistic level. However, the relationship between the parameters cannot be determined compared to the negative correlation yielded by the UD approach (see Figure 14).

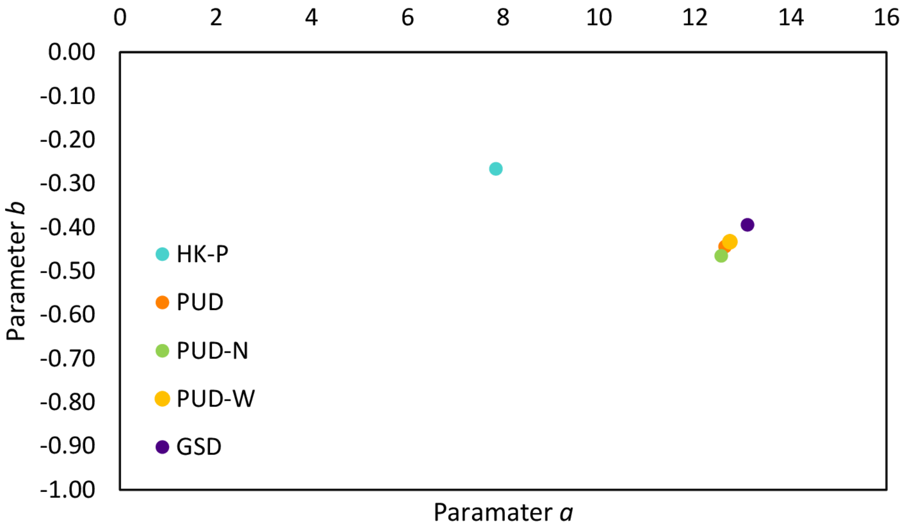


Figure 14. The parameters a and b of M1 for the triplet of the sentence, clause and word – punctuation approach.

Thirdly, the range of the sentence lengths decreased in all the samples, in the case of GSD from $1 \leq SL \leq 17.36$ to $1 \leq SL \leq 8.38$. Despite SLs not exceeding the upper threshold of short-term memory, two issues remain to tackle. While the UD annotation might be overly grained, the segment between two punctuation marks might not exactly correspond to the clause in Chinese (c.f. Chen, 2018; Chen and Liu, 2019, 2022), especially with respect to the fact that Western-based punctuation was integrated into Chinese relatively recently (see Chapter 3.2.2). Another issue arises with the mean clausal lengths of one-clause sentences MCL_1s. Although SLs do not exceed the upper limit of the short-term memory span, MCL_1s do except for HK-P.

The latter issue also appears in GSD, PUD and its versions when the UD approach is applied. It might imply that the word is not a proper measurement unit for the clause with respect to a sufficient granularity of the unit hierarchy. The only exception is HK-P, where not only MCL_1 but also other mean clause lengths are shorter than in the other samples. As mentioned above, HK-P rather represents the spoken text type which generally has shorter clauses.

4.1.2 The sentential phrase and word as constituents

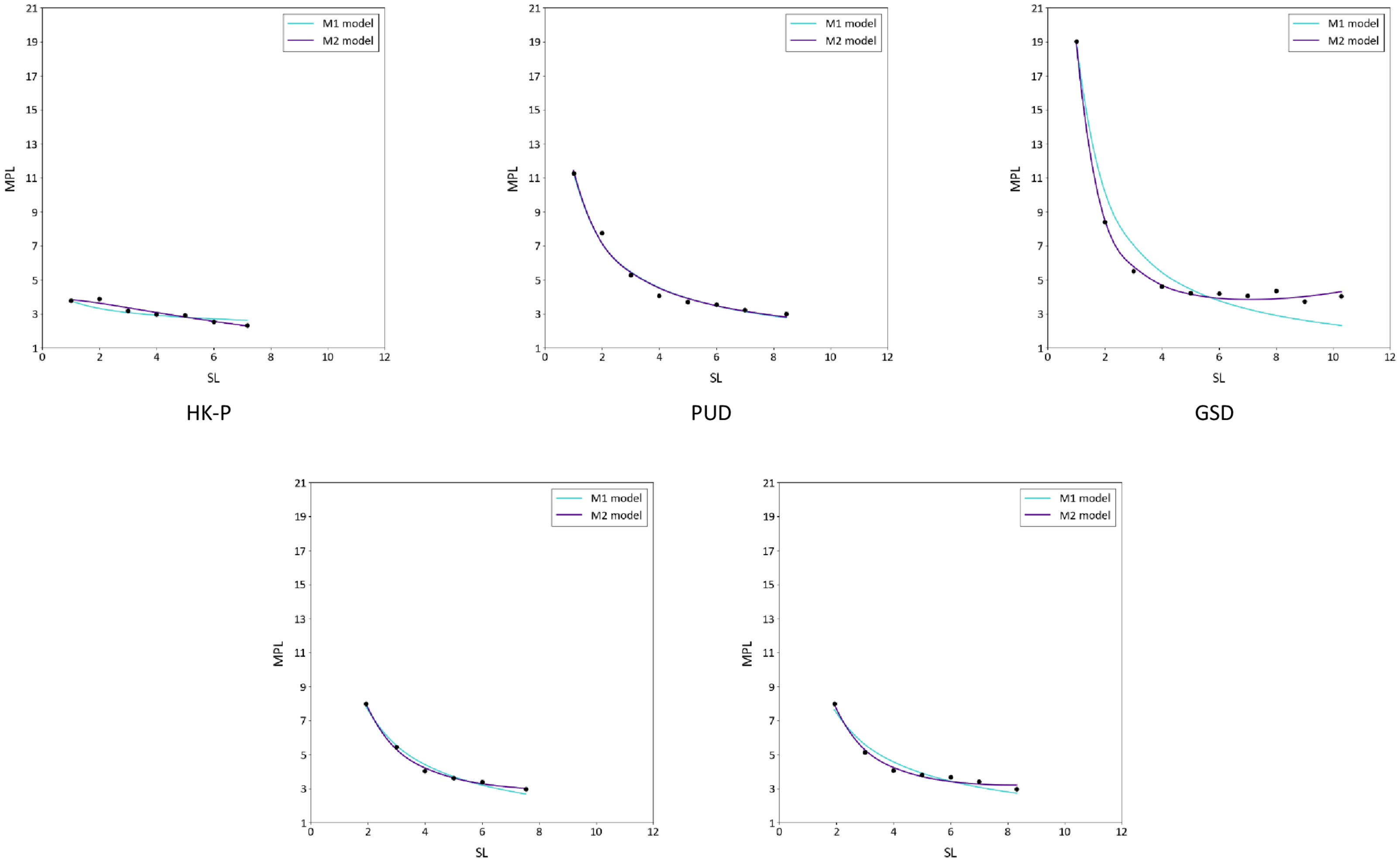
Hypothesis: The longer the sentence length measured in the number of sentential phrases, the shorter the mean length of the sentential phrases measured in words.

Table 17 and Figure 15 show results achieved when the phrase is the direct constituent of the sentence. SL denotes the sentence length measured in the number of phrases, $f(SL)$ its frequency and MPL the mean phrase length measured in the number of words. The data are fitted by both the models, i.e. $y(x) = ax^b$ labelled as M1 and $y(x) = ax^b e^{cx}$ labelled as M2. The values of their parameters (a, b, c) and the coefficient of determination R^2 are presented in the table. As for M1, we use the mean phrase length of one-phrase sentences, i.e. MPL_1, as the parameter a except for PUD-N and PUD-W. In these two samples, one-phrase sentences SL_1 are pooled with the neighbouring construct length SL_2 due to their insufficient frequency ($f(SL_1) = 7$ and $f(SL_1) = 4$ accordingly). Consequently, the pooled SL_1s equal 1.93 (and not 1) and the equation $MPL_1 = aSL_1{}^b = a1^b = a$ is no longer valid (cf. Köhler, 1982, p. 110). For this reason, we exceptionally calculate the value of the parameter a by means of the NLREG software. Finally, if a value of MPL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 17. MAL applied to the triplet of the sentence, phrase and word.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MPL*	*SL*	*f(SL)*	*MPL*	*SL*	*f(SL)*	*MPL*	*SL*	*f(SL)*	*MPL*	*SL*	*f(SL)*	*MPL*
1	13	3.77	1	11	11.27	1.93	102	7.98	1.93	59	7.99	1	40	19.03
2	63	3.87	2	150	7.74	3	120	5.44	3	115	5.14	2	360	8.41
3	82	3.17	3	235	5.29	4	106	4.03	4	126	4.07	3	895	5.51
4	91	2.98	4	232	4.05	5	101	3.63	5	88	3.80	4	1034	4.62
5	62	2.93	5	189	3.71	6	45	3.38	6	65	3.67	5	878	4.21
6	25	2.53	6	110	3.55	7.54	26	2.98	7	31	3.41	6	485	4.21
7.19	16	2.32	7	49	3.24				8.31	16	2.97	7	179	4.07
			8.46	24	2.99							8	80	4.34
												9	32	3.74
												10.29	14	4.04
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	3.77	4.27	*a*	11.27	11.30	*a*	13.16	14.18	*a*	12.13	13.83	*a*	19.03	15.46
b	-0.18	0.07	*b*	-0.66	-0.69	*b*	-0.79	-1.24	*b*	-0.70	-1.30	*b*	-0.90	-1.45
c		0.11	*c*		-0.01	*c*		-0.13	*c*		-0.16	*c*		-0.21
R^2	0.7728	0.9379	R^2	0.9875	0.9879	R^2	0.9825	0.9964	R^2	0.9530	0.9884	R^2	0.9332	0.9970

*calculated by means of the NLREG software



HK-P PUD GSD

PUD-N PUD-W

Figure 15. MAL applied to the triplet of the sentence, phrase and word.

The goodness-of-fit is in accord with the standard of $R^2 \geq 0.90$ and the hypothesis is not rejected except for HK-P fitted by M1. HK-P is the only sample where MPL_2 has the highest value contradicting the menzerathian decreasing tendency (highlighted in yellow in Table 17). The second regime of MPL_2 might not be caused by an unusual behaviour of phrases belonging to two-phrase sentences SL_2, but by phrases of one-phrase sentences SL_1 which are the least frequent. SL_1 makes up only of $3.69\ \%$ of all sentences in HK-P.[116] Seven out of 13 one-phrase sentences are short responses of speakers (e.g. 好的。, *hǎo de*, 'ok; all right'; 不要紧。, *bù yàojǐn*, 'never mind; not important'; 谢谢主席。, *xièxie zhǔxí*, 'thank you, Chairman'). Their phrases consist only of one word (a root is excluded), which considerably lowers MPL_1. The increasing trend in $MPLs$ also appears with longer SLs in GSD (highlighted in yellow in Table 17) – specific properties or a lower frequency of SLs can be taken into account.

As for the M1 parameters, HK-P is not considered because $R^2 < 0.90$. Hence, the parameter a reaches the lowest value in PUD, i.e. $a = 11.27$. GSD shows its highest value which equals 19.03. PUD also has the highest parameter b ($b = -0.66$) and GSD the lowest ($b = -0.90$). The negative correlation between values of both the parameters occurs – the higher the value of the former, the lower the value of the latter (see Figure 16).

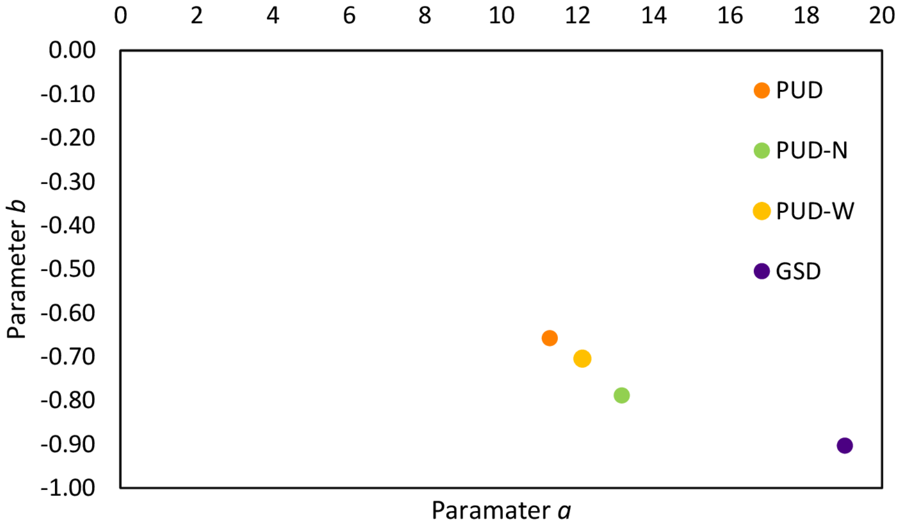


Figure 16. The parameters a and b of M1 for the triplet of the sentence, phrase and word (excluding HK-P).

The issue of the SLs being considerably above the memory span limit (i.e. 7 ± 2, Miller, 1956) does not arise within this triplet. In the case of HK-P, PUD and its versions, the sentence lengths measured in phrases are in a similar range as the sentences measured in clauses. They mainly differ in the frequency distribution. In the case of GSD, not only the frequency distribution but also the scale of SLs considerably changed. While SL had up to 17.36 clauses in the previous triplet, SL of this triplet has only up to 10.29 phrases. It is noteworthy that the

[116] SL_1 made up of $21.19\ \%$ of all sentences in the previous triplet of the sentence, clause and word.

clauses were processed on all levels of depth in dependency trees, whereas the phrases are processed only on the levels immediately neighbouring the root.

As for MPL, the HK-P, PUD-N and PUD-W samples do not struggle with exceeding the upper limit of the short-term memory span (7 ± 2, Miller, 1956), but the upper threshold is exceeded by MPL_1 in PUD and GSD.[117] Firstly, SL_1s have an extremely low frequency. They form $1.10\ \%$ of all sentences in PUD and 1.00% in GSD, which might result in irregular behaviour of their MPL_1s (as in the case of HK-P described above).[118] Secondly, the determination of the phrase as a whole subtree directly depending on a root can contribute to higher MPL_1s. The most frequent construction of SL_1s consists of a root governing a clausal complement (i.e. ccomp).[119] The roots are mostly expressed by the stative verb of existence (有, *yǒu*, 'to be; to exist') or by verb phrases (e.g. 传说, *chuánshuō* or 据称, *jùchēng*, bearing the meaning 'it is said that') and their clausal complements govern complex structures having a high number of words. Thirdly, the UD annotation of some words is inaccurate, resulting in biased syntactic structures. To illustrate the point, we can use an example of the word 因为 (*yīnwèi*, 'because; for; on account of') which is annotated as a root and a verb while being conjunction in the following one-phrase sentence 因为我们不一定能理解和辨识融合了外星思维和高等外星科技的物品。(*Yīnwèi wǒmen bù yī̦dìng néng lǐjiě hé biànshí rónghé le wàixīng sīwéi hé gāoděng wàixīng kējì de wùpǐn.*, 'Because we may not be necessarily able to understand and recognise objects which combine alien thinking and advanced alien technologies.', GSD, sentence ID train-s3359).

Mačutek, Čech and Milička (2017) also yielded a higher value of MPL_1 when analysing the Prague Dependency Treebank 3.0 (Bejček et al., 2013). Despite SL_1 not suffering from an extremely low frequency, "there are 7,125 clauses (more than 12%) which contain only one phrase, and their mean length in words is 9.47 (which means that are many phrases longer than 9.47)" (Mačutek, Čech and Courtin, 2021, p. 2). Their results and the results yielded by the thesis indicate that the high lengths of MPL_1 are caused by coordinate and subordinate clauses which the sentential phrases include (e.g. the root governing ccomp as mentioned above).[120] Hence, the appropriateness of the units chosen for this triplet is brought into question. The phrase does not appear to be the direct measurement unit for the sentence – a linguistic level might be skipped (e.g. a clause).

[117] However, MPL_1 is pooled in PUD-N and PUD-W.

[118] SL_1 of the previous triplet formed $17.50\ \%$ of all sentences in PUD and 10.18% in GSD.

[119] Nine out of 11 sentences have this structure in PUD and 27 out of 40 in GSD.

[120] Seven out of 11 one-phrase sentences in PUD and 35 out of 40 one-phrase sentences in GSD contain at least one clause (i.e. one clausal dependency relation).

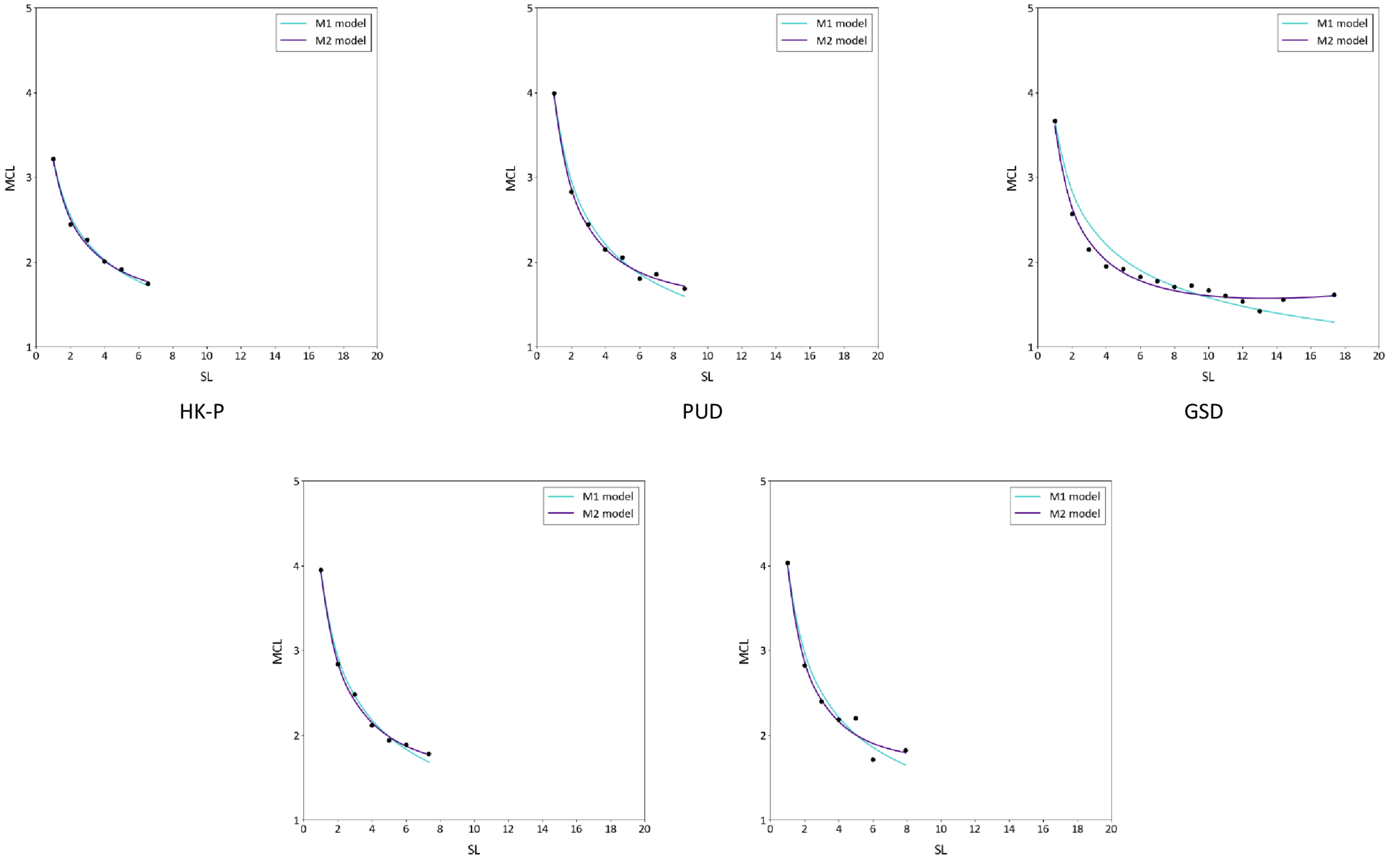
4.1.3 The clause and clausal phrase as constituents

Hypothesis: The longer the sentence length measured in the number of clauses, the shorter the mean length of the clauses measured in clausal phrases.

The results of the clause and the phrase being direct and indirect constituents of the sentence accordingly are presented in Table 18 and Figure 17. SL stands for the sentence length measured in the number of clauses, $f(SL)$ for its frequency and MCL for the mean clause length measured in the number of phrases. The parameters (a , b , c) and the coefficient of determination R^2 of the truncated model M1, i.e. $y(x) = ax^b$, and the complete model M2, i.e. $y(x) = ax^b e^{cx}$, are included in the table. We use MCL_1 as the parameter a to fit M1 to the data. Finally, if a value of MCL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 18. MAL applied to the triplet of the sentence, clause and phrase.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*
1	73	3.22	1	175	3.99	1	83	3.95	1	92	4.03	1	407	3.66
2	97	2.44	2	271	2.83	2	115	2.84	2	156	2.82	2	840	2.57
3	73	2.26	3	248	2.44	3	133	2.48	3	115	2.40	3	830	2.15
4	56	2.01	4	140	2.15	4	77	2.12	4	63	2.19	4	636	1.95
5	23	1.91	5	84	2.06	5	46	1.94	5	38	2.20	5	446	1.92
6.57	30	1.74	6	51	1.81	6	28	1.89	6	23	1.71	6	290	1.83
			7	20	1.86	7.33	18	1.78	7.92	13	1.82	7	186	1.77
			8.64	11	1.69							8	137	1.71
												9	93	1.72
												10	39	1.67
												11	40	1.60
												12	18	1.54
												13	11	1.42
												14.38	13	1.56
												17.36	11	1.61
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	3.22	3.13	*a*	3.99	3.84	*a*	3.95	3.81	*a*	4.03	3.83	*a*	3.66	3.47
b	-0.33	-0.38	*b*	-0.43	-0.52	*b*	-0.43	-0.52	*b*	-0.43	-0.55	*b*	-0.37	-0.50
c		-0.02	*c*		-0.04	*c*		-0.04	*c*		-0.05	*c*		-0.04
R^2	0.9893	0.9931	R^2	0.9865	0.9959	R^2	0.9921	0.9974	R^2	0.9660	0.9786	R^2	0.9069	0.9833



HK-P PUD GSD

PUD-N PUD-W

Figure 17. MAL applied to the triplet of the sentence, clause and phrase.

Making the hierarchy of the language units more granular does not lower the goodness-of-fit. R^2 of both the models meets the standard of $R^2 \geq 0.90$ in all the samples. Hence, the hypothesis is not rejected. The increase in $MCLs$ in PUD, PUD-W and GSD mostly occurs with longer SLs (highlighted in yellow in Table 18) and might be caused either by a low frequency of given SLs or their specific properties. Moreover, SLs in GSD struggles with the same issue of exceeding the upper limit of the short-term memory span (i.e. 7 ± 2, Miller, 1956) as in the case of the triplet composed of the sentence, clause and word (see Chapter 4.1.1).

The parameter a of M1 reaches the lowest value in HK-P ($a = 3.22$) and the highest value in PUD-W ($a = 4.03$). The differences in their values across the samples are not as striking as in the first triplet on the sentence level. Hence, the inclusion of the phrase, or more generally, a linguistic level seems to have a stronger influence than a text type in this case. Neither considerable differences between the parameters b of M1 are observed. Their values range only between -0.33 (HK-P) and -0.43 (PUD-W). Hence, the slope of the curves decreases with a similar 'speed' across the samples. As for the relationship between the parameters, taking HK-P, PUD and GSD into account, the trend of their values being negatively correlated appears (see Figure 18). In the case of PUD and its versions, the differences between the parameters are minimal.

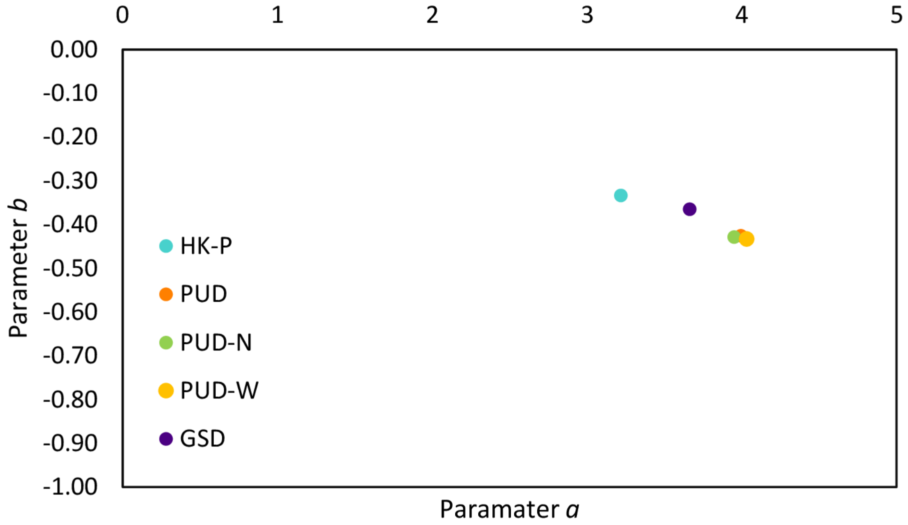
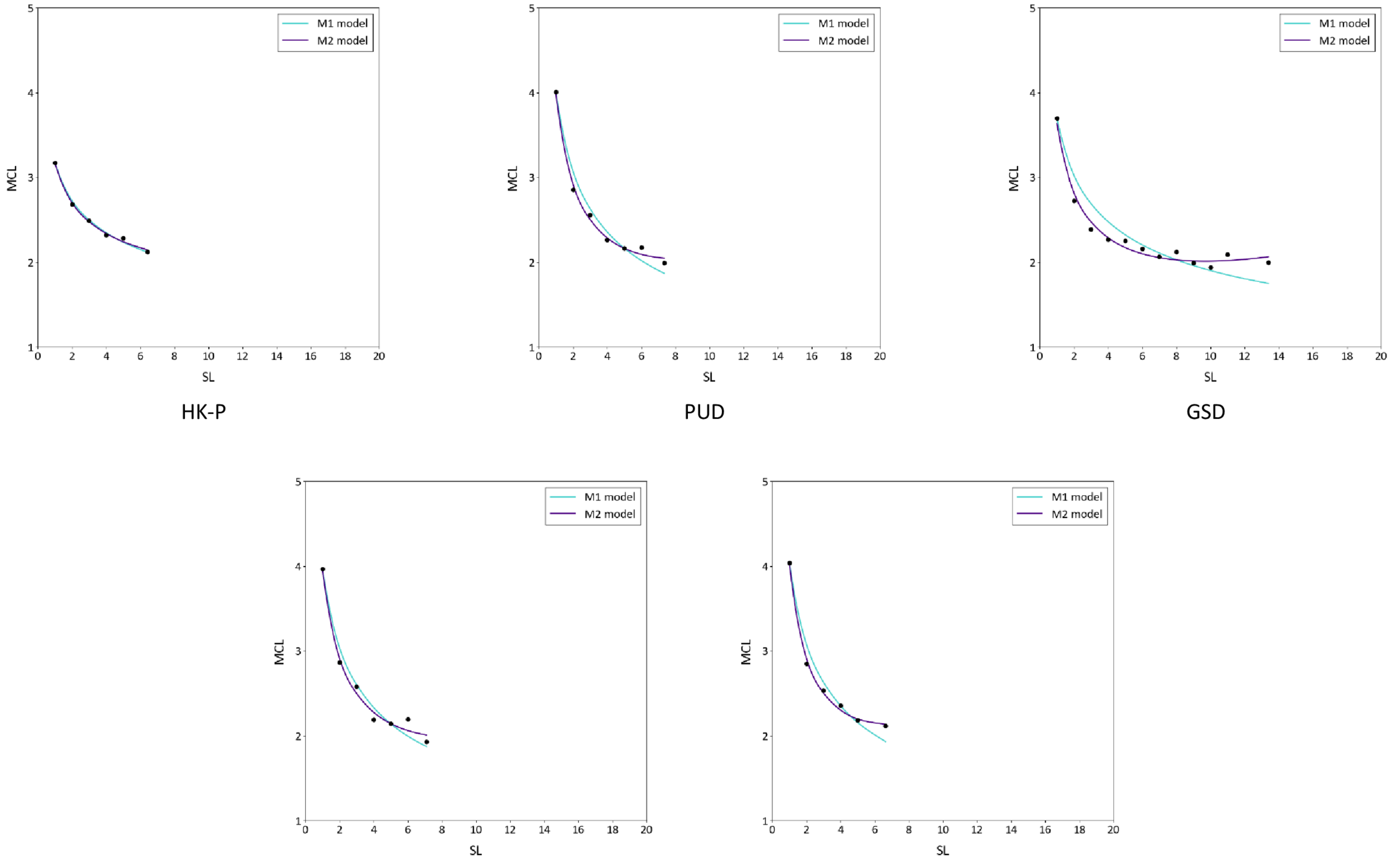


Figure 18. The parameters a and b of M1 for the triplet of the sentence, clause and phrase.

As addressed in Chapter 3.2.3, when processing the clausal phrases, clausal heads are not parts of the phrases and are not the phrases themselves. The exclusion of the clausal heads raises an issue of how to treat a clause consisting only of its head with respect to the sentence length and mean clause length. So far, we have adopted the following approach – if a clause without phrases is identified, it is included in the sum of clauses in a sentence while the number of phrases remains the same. The question is to which extent this approach influences SL and MCL. Hence, we also test an alternative approach – we entirely exclude clauses without phrases from the analysis, i.e. sums of clauses in sentences do not include them. The results are presented in Table 19 and Figure 19.

Table 19. MAL applied to the triplet of the sentence, clause and phrase – the exclusion of clauses without phrases.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*
1	92	3.17	1	192	4.01	1	92	3.97	1	100	4.04	1	593	3.69
2	110	2.68	2	300	2.86	2	127	2.87	2	173	2.85	2	1059	2.72
3	78	2.49	3	252	2.56	3	140	2.58	3	112	2.53	3	858	2.39
4	46	2.32	4	130	2.26	4	74	2.19	4	56	2.36	4	589	2.27
5	14	2.29	5	78	2.16	5	38	2.14	5	40	2.19	5	383	2.25
6.42	12	2.12	6	31	2.17	6	19	2.19	6.63	19	2.12	6	203	2.16
			7.35	17	1.99	7.10	10	1.93				7	132	2.06
												8	86	2.12
												9	40	1.99
												10	21	1.94
												11	18	2.09
												13.40	15	2.00
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	3.17	3.13	*a*	4.01	3.74	*a*	3.97	3.74	*a*	4.04	3.70	*a*	3.69	3.48
b	-0.22	-0.25	*b*	-0.38	-0.54	*b*	-0.38	-0.52	*b*	-0.39	-0.58	*b*	-0.29	-0.43
c		-0.01	*c*		-0.07	*c*		-0.06	*c*		-0.08	*c*		-0.04
R^2	0.9922	0.9950	R^2	0.9654	0.9939	R^2	0.9679	0.9862	R^2	0.9611	0.9964	R^2	0.8583	0.9784



HK-P PUD GSD

PUD-N PUD-W

Figure 19. MAL applied to the triplet of the sentence, clause and phrase – the exclusion of clauses without phrases.

The alternative approach brings about considerable changes only in GSD where, contrary to the previous results, the coefficient of determination R^2 of M1 does not reach the standard of $R^2 \geq 0.90$ and the hypothesis is rejected (i.e. $R^2 = 0.8583$). As for the impact on the unit lengths, SLs only slightly decreased except for GSD (even though its SLs still exceed the upper threshold of the short-term memory span), and $MCLs$ only slightly increased. The parameter a of M1 reaches similar values compared to the previous results, while values of the parameter b of M1 are higher.[121] To sum it up, the overall impact on the results is minimal on this level.

On the one hand, both the approaches face the methodological difficulty in disregarding the clausal heads. On the other hand, measuring the clause in phrases brings a higher granularity to the unit hierarchy. $MCLs$ are in accord with the upper limit of short-term memory in all the samples, whereas most of MCL_1s and MPL_1s measured in words were not. Hence, we again pose both the questions of whether the word is the direct measurement unit of the clause and whether the phrase is the direct measurement unit of the sentence. The results achieved in this triplet lead us to the assumption that the sentence, clause and phrase represent the appropriate unit combination.

[121] Since the results of both approaches do not considerably differ from each other, we do not include a graph displaying the values of both parameters, a and b.

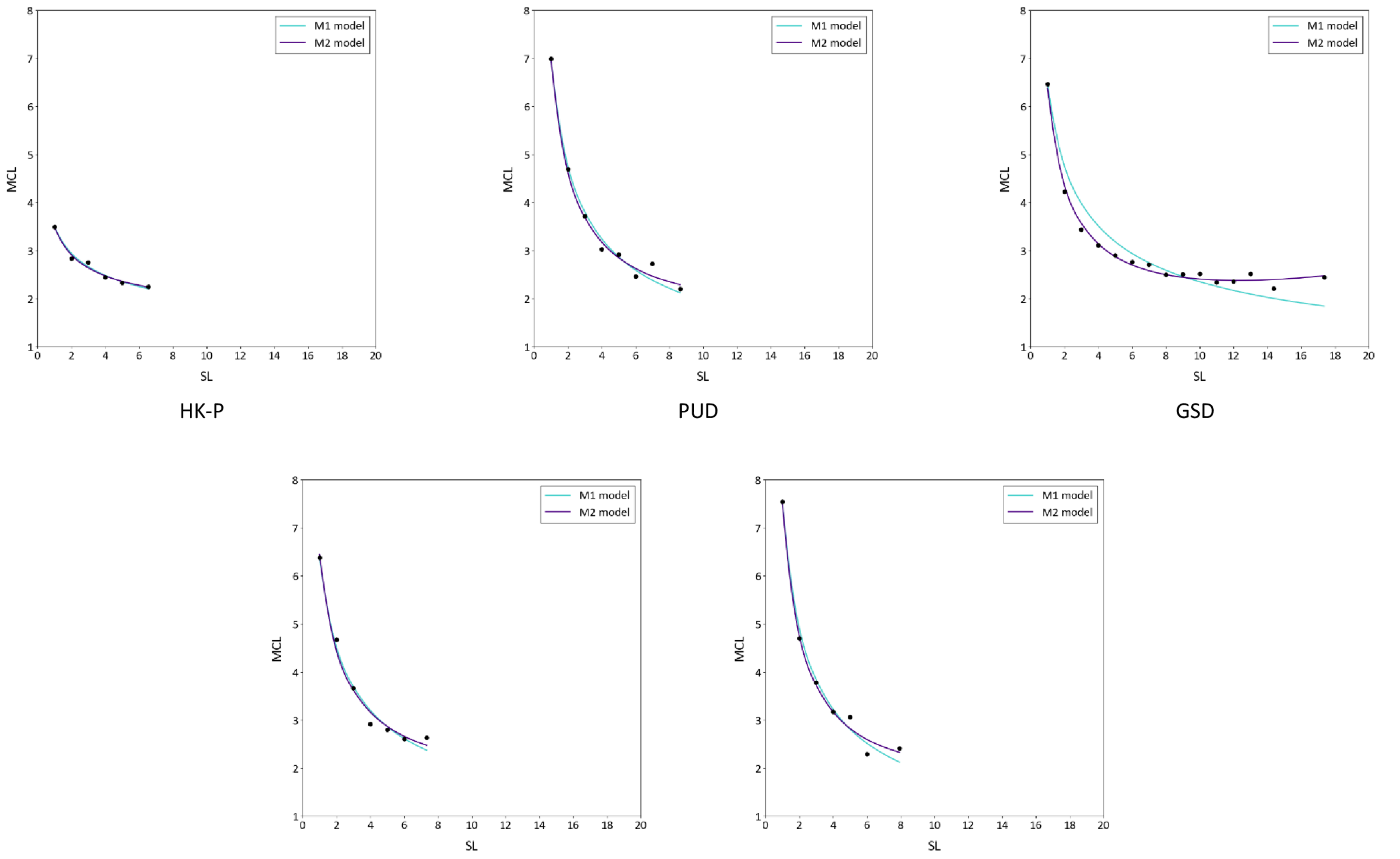
4.1.4 The clause and linear dependency segment as constituents

Hypothesis: 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).

Finally, the results of the last triplet where the role of the indirect constituent is assigned to LDS are presented in Table 20 and Figure 20. SL labels the sentence length measured in the number of clauses, $f(SL)$ its frequency and MCL the mean length of the clause measured in the number of LDSs. The values of the parameters (a, b, c) and the coefficient of determination R^2 of both the models, i.e. $y(x) = ax^b$ with the label M1 and $y(x) = ax^b e^{cx}$ with the label M2, can be found in the table. In the case of M1, we replace the parameter a by MCL_1. Finally, if a value of MCL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 20. MAL applied to the triplet of the sentence, clause and linear dependency segment.

HK-P			PUD			PUD-N			PUD-W			GSD		
SL	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*	*SL*	*f(SL)*	*MCL*
1	75	3.49	1	175	6.99	1	83	6.39	1	92	7.54	1	407	6.46
2	97	2.84	2	271	4.69	2	115	4.67	2	156	4.70	2	840	4.23
3	73	2.75	3	248	3.72	3	133	3.66	3	115	3.78	3	830	3.43
4	56	2.44	4	140	3.03	4	77	2.92	4	63	3.17	4	636	3.10
5	23	2.33	5	84	2.92	5	46	2.80	5	38	3.06	5	446	2.90
6.57	30	2.25	6	51	2.47	6	28	2.61	6	23	2.29	6	290	2.77
			7	20	2.73	7.33	18	2.64	7.92	13	2.41	7	186	2.71
			8.64	11	2.20							8	137	2.50
												9	93	2.51
												10	39	2.52
												11	40	2.34
												12	18	2.35
												13	11	2.52
												14.38	13	2.21
												17.36	11	2.44
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	3.49	3.43	*a*	6.99	6.75	*a*	6.39	6.25	*a*	7.54	7.18	*a*	6.46	6.06
b	-0.25	-0.28	*b*	-0.55	-0.65	*b*	-0.50	-0.58	*b*	-0.61	-0.73	*b*	-0.44	-0.62
c		-0.01	*c*		-0.04	*c*		-0.03	*c*		-0.05	*c*		-0.05
R^2	0.9781	0.9811	R^2	0.9882	0.9924	R^2	0.9847	0.9872	R^2	0.9871	0.9917	R^2	0.9073	0.9916



HK-P PUD GSD

PUD-N PUD-W

Figure 20. MAL applied to the triplet of the sentence, clause and linear dependency segment.

Based on the goodness-of-fit between the models and the data, we can conclude that the hypothesis is corroborated. The standard of $R^2 \geq 0.90$ is reached in all the samples. Nevertheless, the reverse regime is no exception to this last triplet. The increase in $MCLs$ appears with longer SLs of PUD, PUD-N and PUD-W (highlighted in yellow in Table 20) and might be primarily associated with a lower frequency of SLs and an irregular behaviour of their constituents. In the case of GSD, $MCLs$ are affected by the second regime to a greater extent – the regime occurs within SLs in the range of $9 \leq SL \leq 17.36$. However, contrary to the other samples, these SLs suffer not only from a lower frequency but also from their wild scale (see Chapter 4.1.1).[122] Hence, $MCLs$ might also behave irregularly in this case.

Regarding the M1 parameters, the parameter a is the lowest in HK-P ($a = 3.49$) and the highest in PUD-W ($a = 7.54$). HK-P appears to be influenced by both – the phrasal determination (i.e. linguistic level) and text type. The parameter b reaches the lowest value in PUD-W ($b = -0.61$) and the highest value in HK-P ($b = -0.25$). Values of both the parameters, a and b, are negatively correlated across the samples (see Figure 21).

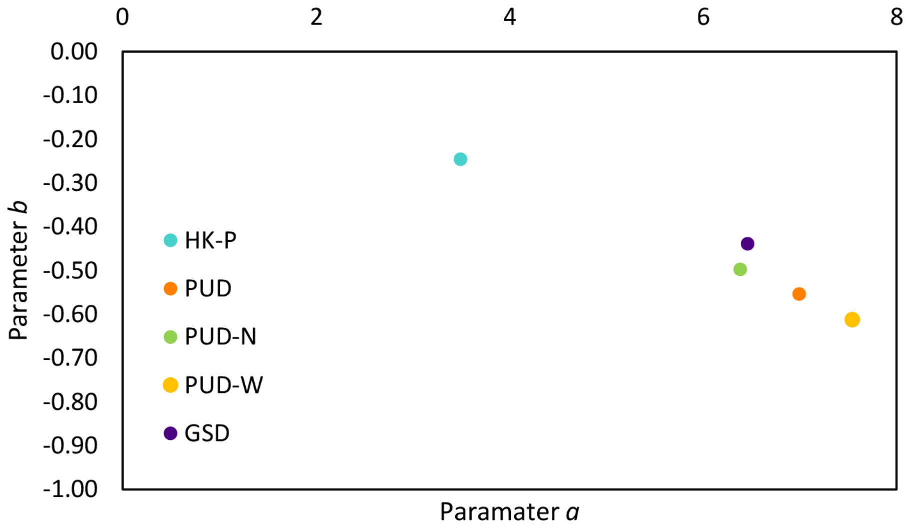


Figure 21. The parameters a and b of M1 for the triplet of the sentence, clause and linear dependency segment.

The issue of SL being above the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) in GSD arises again as in the two previous triplets including the clause as the direct constituent to the sentence (see Chapter 4.1.1 and 4.1.3). Regarding $MCLs$, the change to LDS does not violate the upper threshold, although $MCLs$ increased across the samples in comparison to the previous triplet. The clausal phrase systematically includes elements that depend only on a clausal head and are not clauses themselves. LDS does not take only the dependency syntactic criterion into account but also considers the criterion of the word order, which makes clauses more fragmented. While $MCLs$ measured in the number of clausal phrases were similar in all the samples without regard to a text type under analysis, LDS brings back the

[122] Fluctuating between 0.28 % and 2.33 % of all sentences in GSD.

differences between HK-P and other samples. The lower $MCLs$ of HK-P again indicate the joint influence of the linguistic level (i.e. the phrase determination) and text type, as mentioned above.

Mačutek, Čech and Courtin (2021) tested LDS on the Czech language and Surface Syntactic Universal Dependencies (SUD) treebanks (Gerdes et al., 2018) which annotation differs from UD (for more information, see Chapter 3.2.1). If we compare the results, firstly, their sentence lengths measured in clauses did not exceed the upper limit of the short-term memory as they did in GSD, even though the authors analysed more than 86k sentences, contrary to GSD having only around 4k sentences.[123] However, the authors determined the clause based on the presence of a finite verb. Secondly, regarding the mean clause lengths in LDSs, the scales are similar despite different language material under analysis. Finally, the second regime occurred in their data to a minimal extent and only with extremely low frequent outliers. Since SUD does not directly annotate the clausal dependency relations[124] and we cannot determine the clause based only on finite verbs (e.g. verbs cannot be inflected in Chinese, see Chapter 3.2.2), we cannot test the exact approach adopted by Mačutek, Čech and Courtin (2021).

[123] Mačutek, Čech and Courtin (2021) did not include sentence lengths longer than eight clauses in their analysis. However, the authors excluded the sentences not because of their lengths but because of their relative frequency lower than $0.10\ \%$. If we apply this condition to GSD, none of SLs would be excluded – each SL has a frequency higher than $0.10\ \%$.

[124] The SUD standard annotation does not distinguish between the nominal clausal subject (`nsubj`) and the clausal subject (`csubj`) used by UD. Hence, they are annotated as a single dependency relation, i.e. `subj`. Similarly, the clausal complement (`ccomp`) and the open clausal complement (`xcomp`) are annotated as `comp:obj` in SUD, and the adverbial clause modifier (`advcl`) and the clausal modifier of a noun (`acl`) as a single dependency relation `mod` (Gerdes et al., 2018).

4.1.5 Summary of triplets on the sentence level

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 (highlighted in grey in Table 21).

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.

When comparing the clausal phrase and linear dependency segment, we cannot unambiguously conclude based on the goodness-of-fit which unit achieves better results (see Table 21). 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.

As for the parameters (Table 21), linguistic levels involved in the triplets, or more precisely, their determination exerts a strong influence on values of the parameters a. The triplets including the clause and the phrase measured in words yield higher values than the triplets including the clause measured in clausal phrases and linear dependency segments. The parameter a also differs across the samples. It reaches the lowest values in HK-P and the highest values mostly in PUD-W. While HK-P rather represents the spoken language (i.e. proceedings), the rest of the samples represent the written language (news and/or Wikipedia articles). Hence, the text type also influences the parameter a. In the case of the parameter b, a similar trend can be observed – HK-P shows the highest values, whereas PUD-W the lowest. Finally, the values of both the parameters are mostly negatively correlated.

Table 21. The parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained on the sentence level.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	sentence-clause-word									
a	5.63	5.53	12.50	12.05	10.99	10.92	13.87	13.14	11.64	10.91
b	-0.25	-0.29	-0.59	-0.70	-0.52	-0.57	-0.66	-0.79	-0.48	-0.65
c		-0.02		-0.04		-0.02		-0.05		-0.05
R^2	0.9744	0.9779	0.9924	0.9964	0.9890	0.9897	0.9878	0.9923	0.9228	0.9915
	sentence-clause-word – punctuation approach									
a	7.85	7.45	12.63	11.58	12.55	11.28	12.73	11.21	13.10	12.05
b	-0.27	-0.36	-0.44	-0.63	-0.47	-0.68	-0.43	-0.70	-0.39	-0.57
c		-0.05		-0.09		-0.11		-0.13		-0.07
R^2	0.9752	0.9850	0.9793	0.9946	0.9816	0.9927	0.9653	0.9982	0.9257	0.9734
	sentence-phrase-word									
a	3.77	4.27	11.27	11.30	13.16	14.18	12.13	13.83	19.03	15.46
b	-0.18	0.07	-0.66	-0.69	-0.79	-1.24	-0.70	-1.30	-0.90	-1.45
c		0.11		-0.01		-0.13		-0.16		-0.21
R^2	0.7728	0.9379	0.9875	0.9879	0.9825	0.9964	0.9530	0.9884	0.9332	0.9970
	sentence-clause-phrase									
a	3.22	3.13	3.99	3.84	3.95	3.81	4.03	3.83	3.66	3.47
b	-0.33	-0.38	-0.43	-0.52	-0.43	-0.52	-0.43	-0.55	-0.37	-0.50
c		-0.02		-0.04		-0.04		-0.05		-0.04
R^2	0.9893	0.9931	0.9865	0.9959	0.9921	0.9974	0.9660	0.9786	0.9069	0.9833
	sentence-clause-phrase – exclusion of clauses with zero phrases									
a	3.17	3.13	4.01	3.74	3.97	3.74	4.04	3.70	3.69	3.48
b	-0.22	-0.25	-0.38	-0.54	-0.38	-0.52	-0.39	-0.58	-0.29	-0.43
c		-0.01		-0.07		-0.06		-0.08		-0.04
R^2	0.9922	0.9950	0.9654	0.9939	0.9679	0.9862	0.9611	0.9964	0.8583	0.9784

	sentence-clause-linear dependency segment									
a	3.49	3.43	6.99	6.75	6.39	6.25	7.54	7.18	6.46	6.06
b	-0.25	-0.28	-0.55	-0.65	-0.50	-0.58	-0.61	-0.73	-0.44	-0.62
c		-0.01		-0.04		-0.03		-0.05		-0.05
R^2	0.9781	0.9811	0.9882	0.9924	0.9847	0.9872	0.9871	0.9917	0.9073	0.9916

4.2 The clause as the construct

4.2.1 The word and character as constituents

Hypothesis: the longer the clause length measured in the number of words, the shorter the mean length of the words measured in (Chinese) characters[125].

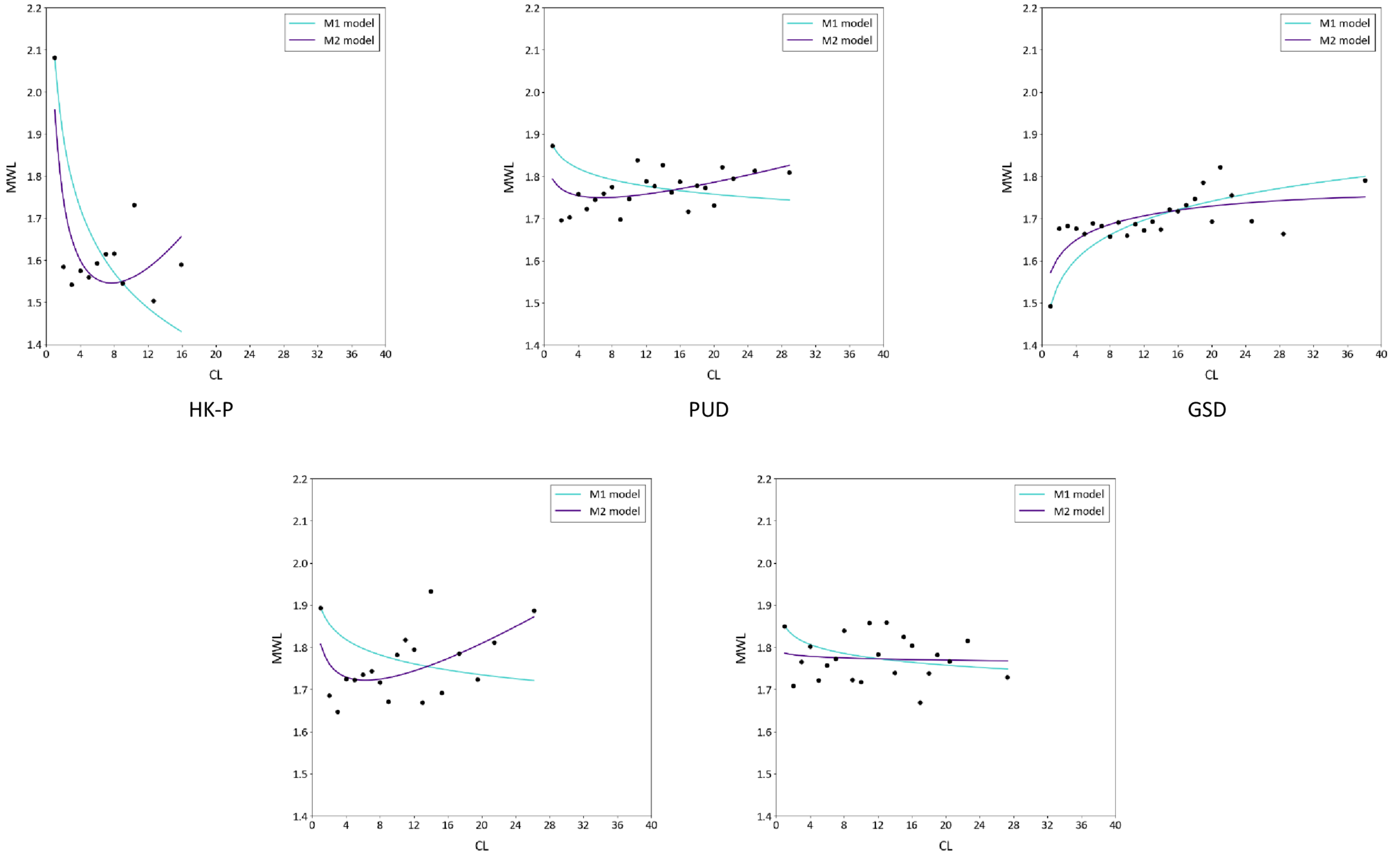
Table 22 and Figure 22 summarise the results yielded on the clause level. CL labels the clause length measured in the number of words, $f(CL)$ its frequency and MWL the mean word length measured in the number of (Chinese) characters. We apply both the models to the data – the truncated model $y(x) = ax^b$ labelled as M1 and the complete model $y(x) = ax^b e^{cx}$ labelled as M2. Their parameters (a, b, c) and coefficient of determination R^2 are included in the table. In the case of M1, the parameter a equals the mean word length of one-word clauses, i.e. MWL_1.

[125] We remind the reader that one Chinese character corresponds to one syllable except for erization.

Table 22. MAL applied to the triplet of the clause, word and character.

HK-P			PUD			PUD-N			PUD-W			GSD		
CL	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*
1	147	2.08	1	213	1.87	1	113	1.89	1	100	1.85	1	2538	1.49
2	159	1.58	2	419	1.70	2	232	1.69	2	187	1.71	2	1878	1.68
3	177	1.54	3	448	1.70	3	235	1.65	3	213	1.77	3	1998	1.68
4	168	1.58	4	374	1.76	4	210	1.73	4	164	1.80	4	1856	1.68
5	109	1.56	5	292	1.72	5	168	1.72	5	124	1.72	5	1490	1.66
6	86	1.59	6	210	1.75	6	114	1.74	6	96	1.76	6	1306	1.69
7	62	1.62	7	188	1.76	7	91	1.74	7	97	1.77	7	1008	1.68
8	41	1.62	8	142	1.77	8	75	1.72	8	67	1.84	8	819	1.66
9	31	1.54	9	101	1.70	9	48	1.67	9	53	1.72	9	712	1.69
10.36	22	1.73	10	101	1.75	10	45	1.78	10	56	1.72	10	512	1.66
12.64	11	1.50	11	85	1.84	11	42	1.82	11	43	1.86	11	394	1.69
15.91	11	1.59	12	77	1.79	12	39	1.79	12	38	1.78	12	391	1.67
			13	60	1.78	13	26	1.67	13	34	1.86	13	250	1.69
			14	40	1.83	14	18	1.93	14	22	1.74	14	198	1.67
			15	51	1.76	15.28	29	1.69	15	30	1.82	15	140	1.72
			16	30	1.79	17.33	24	1.79	16	22	1.80	16	96	1.72
			17	32	1.72	19.53	17	1.72	17	16	1.67	17	89	1.73
			18	18	1.78	21.45	11	1.81	18	10	1.74	18	53	1.75
			19	22	1.77	26.15	13	1.89	19	14	1.78	19	49	1.79
			20	21	1.73				20.43	21	1.77	20	35	1.69

			21	15	1.82				22.57	14	1.82	21	16	1.82
			22.30	20	1.80				27.27	11	1.73	22.36	22	1.76
			24.82	11	1.81							24.70	20	1.69
			28.92	12	1.81							28.46	13	1.66
												38.10	10	1.79
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.08	1.91	*a*	1.87	1.79	*a*	1.89	1.79	*a*	1.85	1.79	*a*	1.49	1.57
b	-0.14	-0.20	*b*	-0.02	-0.02	*b*	-0.03	-0.05	*b*	-0.02	-0.003	*b*	0.05	0.04
c		-0.03	*c*		-0.003	*c*		-0.01	*c*		0.00	*c*		0.001
R^2	NA	0.5862	R^2	NA	0.1936	R^2	NA	0.2602	R^2	NA	0.0076	R^2	0.2368	0.5188



HK-P PUD GSD

PUD-N PUD-W

Figure 22. MAL applied to the triplet of the clause, word and character.

When evaluating the results based on the coefficient of determination R^2 and the standard of $R^2 \geq 0.90$, we can conclude that the goodness-of-fit between the models and the data is extremely unsatisfactory and the samples do not corroborate the hypothesis.

The clause measured in the number of words suffers from the wide scale of its lengths that extensively exceed the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) in each sample. Such results support our assumption which we made on the sentence level that the word is not the direct constituent of the clause.

On the contrary, the scale of $MWLs$ is narrow. The word lengths fluctuate only between one and two (Chinese) characters on average. In general, one- and two-character words prevail in modern Chinese (Chen, Liang and Liu, 2015, p. 8) and this prevalence is confirmed in our samples for both – tokens and types – with only one exception (three-character word types in PUD, see

Figure 23).[126] The question arises whether a construct (in our current case, the clause) can influence the mean lengths of Chinese words being its constituent. The specificity of Chinese in one- and two-character words vastly outweighing other word lengths might not provide the law with enough 'space' to come into play. Or in other words, this specificity might be the boundary condition for the law when the Chinese word becomes the constituent.

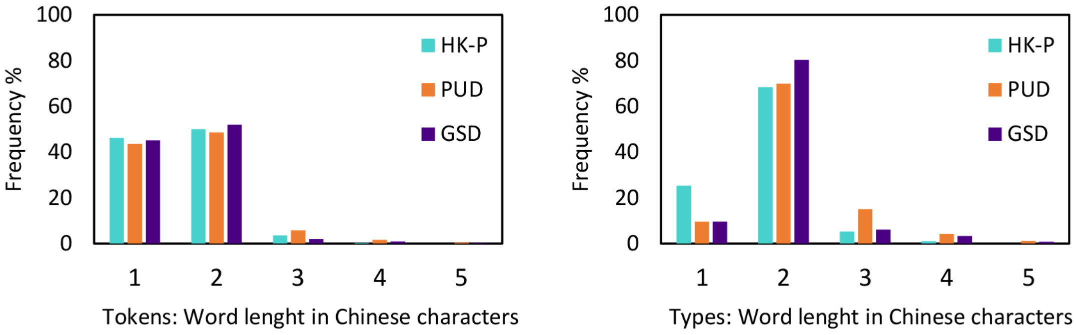


Figure 23. The word length distribution of word tokens and types.

[126] In Figure 23, we include only the word lengths in the range of one to five Chinese characters which make up 99 % of all words (tokens and types) in each sample. Words containing non-Chinese graphemes are excluded.

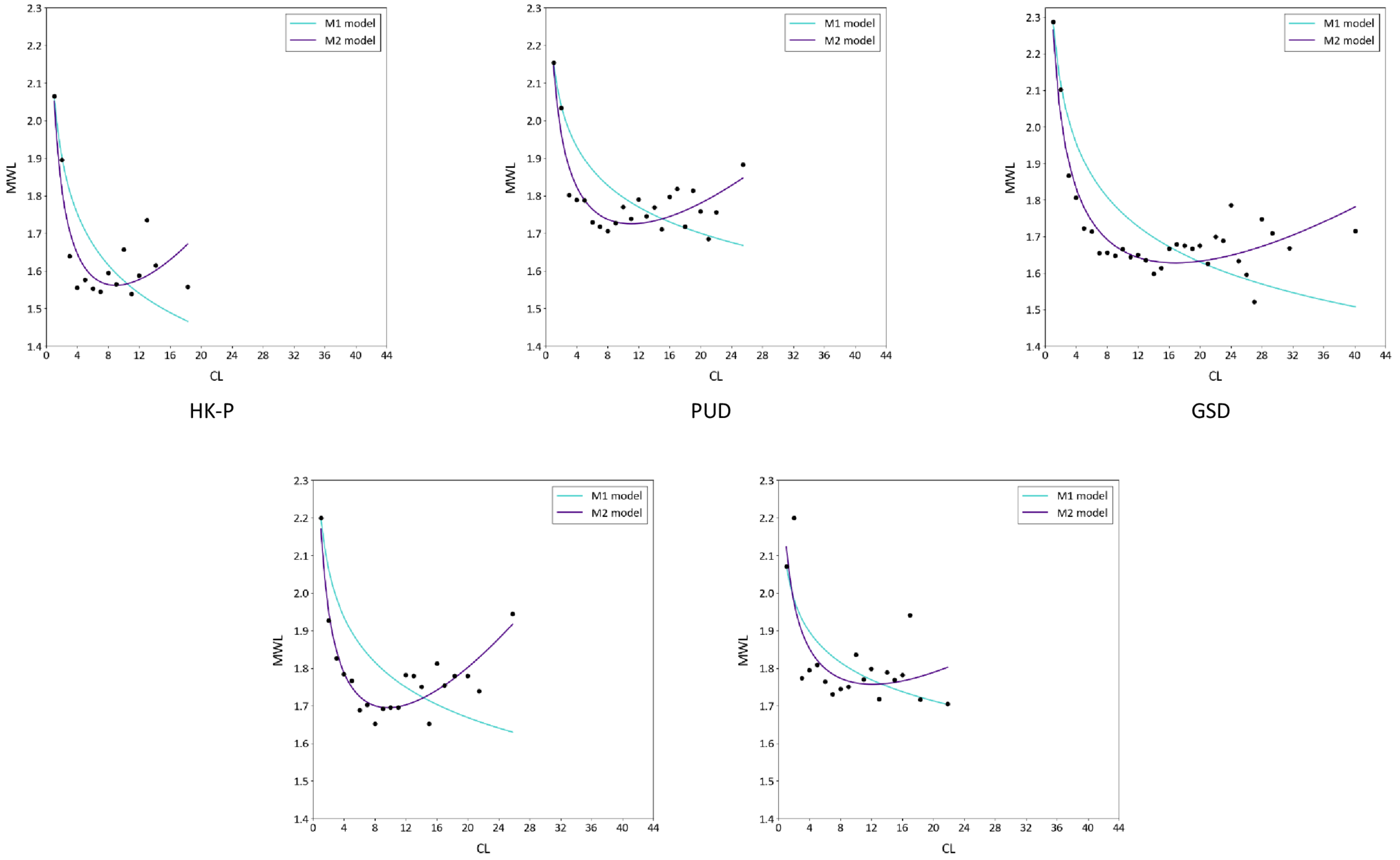
The high variance of CLs in words and the narrow scale of $MWLs$ in (Chinese) characters also appear in other studies (Bohn, 1998, 2002; Hou et al., 2019b; Chen and Liu, 2022). Contrary to our results, the studies yielded the menzerathian decreasing trend of $MWLs$ and the coefficient of determination R^2 reaching the value of 0.70 (Bohn, 1998, 2002; Hou et al., 2019b; Chen and Liu, 2022), the value of 0.80 (Bohn, 1998, 2002) or meeting our standard of $R^2 \geq 0.90$ (Hou et al., 2019b). These results were achieved when the clause was determined as a segment between selected punctuation marks. Hence, the question arises whether the UD annotation of the clausal dependency relations contributes to or causes our unsatisfactory results. Similarly to the sentence level, we apply the punctuation approach to HK-P, PUD (and its versions) and GSD.[127] The results are shown in Table 23 and Figure 24.

[127] We opted for the same punctuation marks as on the sentence level, i.e. a comma '，', a colon '：', a semicolon '；', and an ellipsis '…', '……'. The studies mentioned above agree on the clause determination based on the punctuation marks. However, there is no consensus about the selection. Chen and Liu (2022) selected a comma, Chen and Liu (2019) added the semicolon and Bohn (1998, 2002) and Hou et al. (2019a, 2019b) also used the colon.

Table 23. MAL applied to the triplet of the clause, word and character – punctuation approach.

HK-P			PUD			PUD-N			PUD-W			GSD		
CL	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*	*CL*	*f(CL)*	*MWL*
1	62	2.06	1	78	2.15	1	50	2.20	1	28	2.07	1	202	2.29
2	24	1.90	2	90	2.03	2	55	1.93	2	35	2.20	2	482	2.10
3	37	1.64	3	135	1.80	3	73	1.83	3	62	1.77	3	664	1.87
4	63	1.56	4	158	1.79	4	80	1.78	4	78	1.79	4	906	1.81
5	87	1.58	5	174	1.79	5	84	1.77	5	90	1.81	5	955	1.72
6	75	1.55	6	181	1.73	6	84	1.69	6	97	1.76	6	1126	1.71
7	75	1.54	7	188	1.72	7	85	1.70	7	103	1.73	7	907	1.65
8	46	1.60	8	189	1.71	8	81	1.65	8	108	1.75	8	902	1.66
9	37	1.56	9	176	1.73	9	70	1.69	9	106	1.75	9	759	1.65
10	44	1.66	10	141	1.77	10	65	1.70	10	76	1.84	10	702	1.67
11	28	1.54	11	146	1.74	11	61	1.70	11	85	1.77	11	541	1.64
12	19	1.59	12	125	1.79	12	61	1.78	12	64	1.80	12	466	1.65
13	25	1.74	13	95	1.75	13	43	1.78	13	52	1.72	13	401	1.64
14.13	16	1.61	14	77	1.77	14	41	1.75	14	36	1.79	14	279	1.60
18.27	11	1.56	15	56	1.71	15	28	1.65	15	28	1.77	15	238	1.61
			16	43	1.80	16	22	1.81	16	21	1.78	16	186	1.67
			17	35	1.82	17	23	1.75	17	12	1.94	17	108	1.68
			18	29	1.72	18.29	21	1.78	18.33	21	1.72	18	110	1.68
			19	13	1.81	20	10	1.78	21.86	14	1.71	19	95	1.67
			20	16	1.76	21.47	17	1.74				20	60	1.68

			21	10	1.69	25.80	10	1.94				21	56	1.63
			22	11	1.76							22	31	1.70
			25.50	14	1.88							23	20	1.69
												24	14	1.79
												25	19	1.63
												26	10	1.60
												27	13	1.52
												28	15	1.75
												29.36	11	1.71
												31.55	11	1.67
												40.10	10	1.72
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.06	2.00	*a*	2.15	2.12	*a*	2.20	2.13	*a*	2.07	2.10	*a*	2.29	2.24
b	-0.12	-0.21	*b*	-0.08	-0.15	*b*	-0.09	-0.18	*b*	-0.06	-0.12	*b*	-0.11	-0.18
c		-0.02	*c*		-0.01	*c*		-0.02	*c*		-0.01	*c*		-0.01
R^2	0.4084	0.7812	R^2	0.1903	0.8047	R^2	NA	0.8715	R^2	0.4299	0.5473	R^2	0.3899	0.8593



HK-P PUD GSD

PUD-N PUD-W

Figure 24. MAL applied to the triplet of the clause, word and character – punctuation approach.

The alternative approach to the clause does not corroborate the hypothesis with respect to the standard of $R^2 \geq 0.90$. The considerably better fit, i.e. $0.7812 \leq R^2 \leq 0.8715$, which approximates the results yielded by the studies mentioned above, is only achieved when we fit M2 to HK-P, PUD, PUD-N and GSD. It should be noted that Hou et al. (2019b) and Chen and Liu (2022) also fitted the data with the complete model, which generally gives a better fit at the cost of one extra parameter. On the other hand, the menzerathian decreasing trend is at least indicated in all the samples, which mainly applies to $MWLs$ belonging to clause lengths in the range of $1 \leq CL \leq 8$. The rest of $MWLs$ is heavily affected by the second regime and fluctuates to a higher degree than in the studies mentioned above. Despite the better results brought by the alternative approach, the high variance in the clause lengths (e.g. $1 \leq CL \leq 40.10$ in GSD) and the low variance in the word lengths (between one and two characters on average) remain the issues to tackle. The former supports the assumption of the word not being the direct constituent of the clause, and the latter confirms the specific word length distribution in Chinese.

Finally, Chinese texts usually contain words fully or partly consisting of non-Chinese characters (or graphemes), which might have an impact on a degree of sample homogeneity. While one Chinese grapheme, i.e. Chinese character, roughly corresponds to one syllable, one non-Chinese grapheme usually represents a letter, numeral, or symbol. Studies on Chinese which use the word measured in Chinese characters as the constituent of the clause do not address this issue (Bohn, 1998, 2002; Hou et al., 2019b; Chen and Liu, 2022). Only Berdicevskis (2021) mentions the exclusion of sentences which contain words annotated as symbols (`SYM`) or unidentifiable tokens (`X`). These two categories might capture non-Chinese elements (e.g. Uniform Resource Locator – URL) but do not systematically target words which are fully or partly composed of graphemes not originating from an analysed language.

As for our samples, the non-Chinese words or words mixing both the types of graphemes account for 3.60 % of all word tokens in PUD, 3.03 % in GSD and their proportion in HK-P is the lowest – they do not exceed 0.2 %. Even though we do not expect these words to have a considerable impact on the results, we exclude all clauses which contain at least one non-Chinese grapheme from PUD and GSD and run the analysis again. Table 24 presents the results. As expected, the exclusion does not considerably change the initial results. The goodness-of-fit between the models and the data remains unsatisfactory and the hypothesis remains rejected.

Table 24. The number of clauses $n(C)$, the parameters (a, b, c) and the coefficients of determination R^2 of both the model (M1, M2) obtained by the inclusion and exclusion of clauses containing non-Chinese graphemes.

	PUD				GSD			
	M1	M2	M1	M2	M1	M2	M1	M2
	incl.		excl.		incl.		excl.	
n(C)	2982		2561		15893		13465	
a	1.87	1.79	1.87	1.80	1.49	1.57	1.49	1.58
b	-0.02	-0.02	-0.04	-0.06	0.05	0.04	0.03	-0.01
c		-0.003		-0.01		0.001		-0.002
R^2	NA	0.1936	NA	0.3822	0.2368	0.5188	NA	0.2250

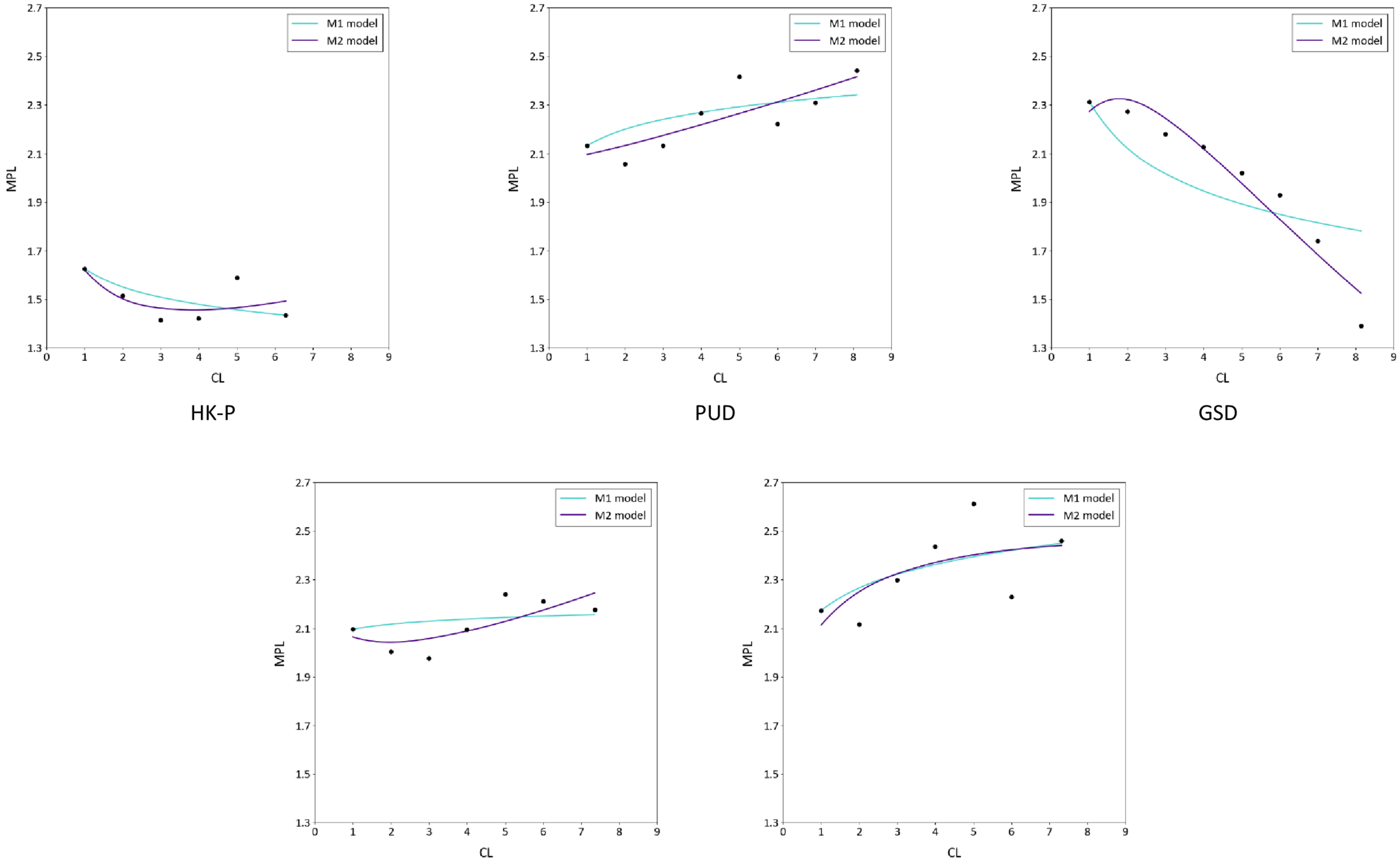
4.2.2 The phrase and word as constituents

Hypothesis: the longer the clause length measured in the number of clausal phrases, the shorter the mean length of the phrases measured in words.

The results obtained when the clause is measured directly in phrases and indirectly in words are presented in Table 25 and Figure 25. CL denotes the clause length measured the in the number of phrases, $f(CL)$ its frequency and MPL the mean phrase length measured in the number of words. The data are fitted by both the models – M1 denoting the truncated model $y(x) = ax^b$ and M2 denoting the complete model $y(x) = ax^b e^{cx}$. Their parameters (a, b, c) and the coefficient of determination R^2 can be found in the table. When fitting M1 to the data, we use the phrase length of the mono-phrasal clauses, i.e. MPL_1, as the parameter a. Finally, if a value of MPL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 25. MAL applied to the triplet of the clause, phrase and word.

HK-P			PUD			PUD-N			PUD-W			GSD		
CL	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*
1	235	1.63	1	730	2.13	1	384	2.10	1	346	2.17	1	3783	2.31
2	251	1.51	2	856	2.06	2	457	2.00	2	399	2.12	2	4273	2.27
3	188	1.41	3	544	2.13	3	279	1.98	3	265	2.30	3	2985	2.18
4	128	1.42	4	343	2.26	4	172	2.09	4	171	2.44	4	1462	2.13
5	54	1.59	5	178	2.42	5	94	2.24	5	84	2.61	5	622	2.02
6.29	21	1.44	6	82	2.22	6	37	2.21	6	45	2.23	6	171	1.93
			7	25	2.31	7.36	14	2.18	7.32	22	2.46	7	45	1.74
			8.09	11	2.44							8.14	14	1.39
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.63	1.55	*a*	2.13	2.05	*a*	2.10	2.00	*a*	2.17	2.14	*a*	2.31	2.54
b	-0.07	-0.18	*b*	0.04	-0.01	*b*	0.01	-0.06	*b*	0.06	0.11	*b*	-0.12	0.19
c		-0.05	*c*		-0.02	*c*		-0.03	*c*		0.01	*c*		0.11
R^2	0.2555	0.4628	R^2	0.5078	0.6592	R^2	0.1598	0.5448	R^2	0.4049	0.4260	R^2	0.6044	0.9385



HK-P PUD GSD

PUD-N PUD-W

Figure 25. MAL applied to the triplet of the clause, phrase and word.

The hypothesis is not corroborated except for GSD fitted by M2 ($R^2 = 0.9385$). GSD is the only sample where $MPLs$ decrease, while $MPLs$ in HK-P and PUD are affected by the second regime (highlighted in yellow in Table 25). The decreasing trend in HK-P is mainly violated by MPL_5 and $MPLs$ in PUD and its versions rather increase (their parameter b of M1 has positive values).

Although we obtained unsatisfactory results, the variance in CLs is reduced with the phrase being the measurement unit of the clause and the upper limit of the short-term memory span (7 ± 2, Miller, 1956) is respected in all the samples. The reduction in the scales of the clause lengths supports our assumption that the word is not the direct constituent of the clause.

As for the constituent, the phrases measured in words also fluctuate in a narrow range, i.e. around two words on average. If we look into the phrase lengths distribution, one-word phrases make up of 75.62 % of all phrases in HK-P, 62.30 % in PUD and 56.70 % in GSD. This prevalence lowers $MPLs$ and reduces differences in their mean lengths. The question arises of what causes the phrases to be so short, especially with respect to PUD and GSD representing the descriptive and informative text type.

Firstly, the governance of the dependency relations in UD should be considered. The prioritisation of the content words tends to arrange tree nodes horizontally than vertically, which flattens syntactic structures (and UD trees). Hence, the clauses incline to consist of more phrases shorter in words rather than vice versa.

Secondly, the results of the previous triplet showed that the punctuation approach reduced the number of clauses compared to the UD approach, i.e. from $1{,}024$ to 649 in HK-P, from $2{,}982$ to $2{,}180$ in PUD and from $15{,}893$ to $10{,}299$ in GSD. The reduction indicates that the phrases are distributed among a higher number of clauses in UD which contributes to lower $MPLs$.

Finally, the determination of the phrases leads to the exclusion of their governors, or more precisely, words functioning as clausal heads. They are neither part of the phrases nor the phrases themselves. The proportion of these excluded words reaches almost 24 % in HK-P and around $15 - 18$ % in PUD and GSD (see Table 26).

Table 26. The numbers of clauses $n(C)$ and words $n(W)$ based on the different approaches to clausal heads.

	all clauses		all clausal heads excl.				clausal heads without phrases excl.			
	N(C)	*N(W)*	*N(C)*	*%*	*N(W)*	*%*	*N(C)*	*%*	*N(W)*	*%*
HK-P	1024	4312	877	85.64	3288	**76.25**	877	85.64	4165	**96.59**
PUD	2982	18513	2769	92.86	15531	**83.89**	2769	92.86	18300	**98.85**
PUD-N	1550	9052	1437	92.71	7502	**82.88**	1437	92.71	8939	**98.75**
PUD-W	1432	9461	1332	93.02	8029	**84.86**	1332	93.02	9361	**98.94**
GSD	15893	84898	13355	84.03	69096	**81.39**	13355	84.03	82451	**97.12**

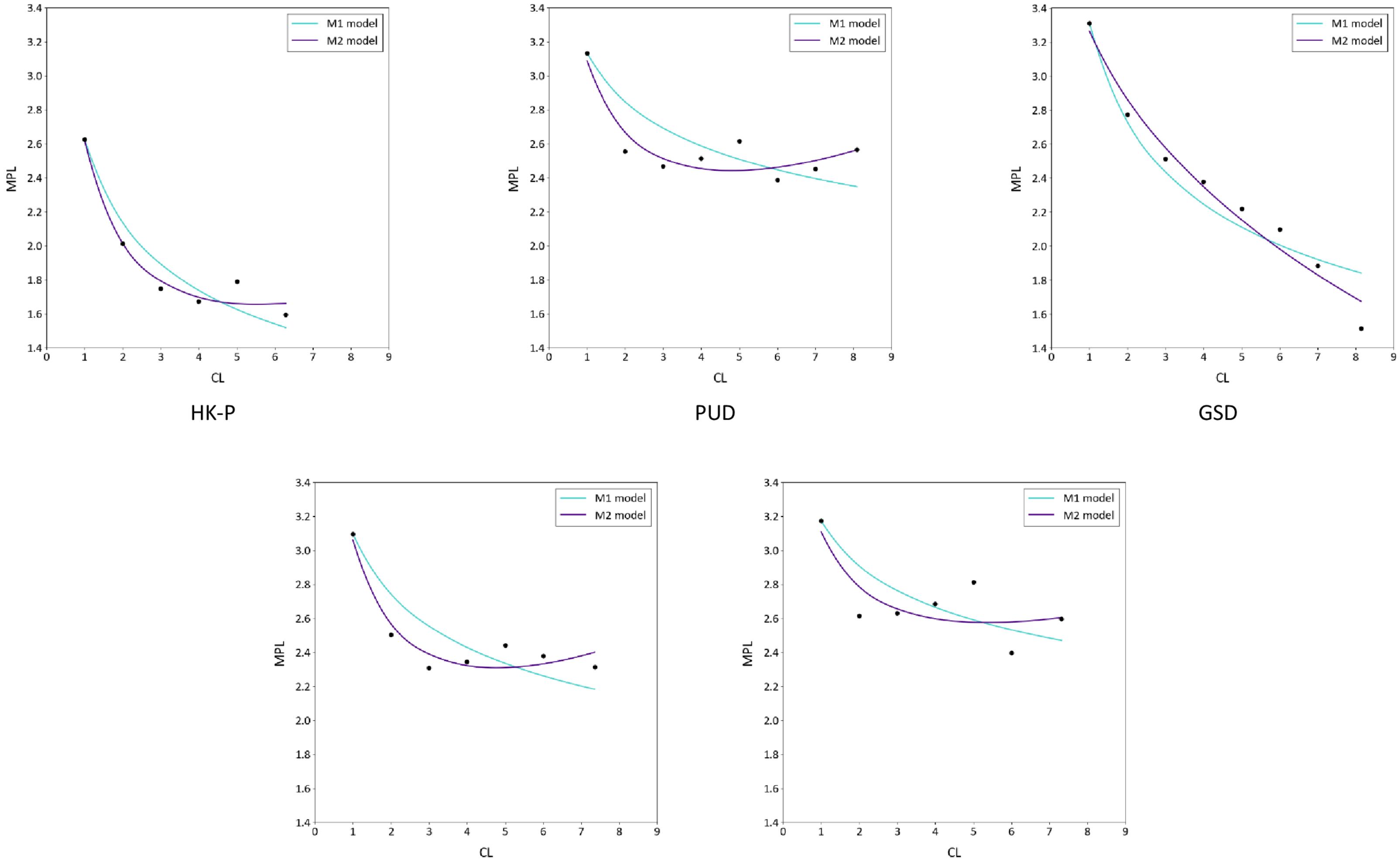
To demonstrate the impact of the phrase determination on $MPLs$, we opt for an alternative approach. Each MPL is calculated using the equation

$$MPL_i = \frac{\sum W_i}{\sum C_i \times i},$$

where i is a given length, MPL_i is the mean phrase length belonging to clauses of the given length i, $\sum W_i$ is the sum of words belonging to the clauses of the length i and $\sum C_i$ is the sum of the clauses of the length i (Chen and Liu, 2022). Using this equation, we include all words which function as clausal heads and govern at least one phrase into $\sum W_i$. The results are presented in Table 27 and Figure 26.

Table 27. MAL applied to the triplet of the clause, phrase and word – inclusion of clausal heads governing at least one phrase.

HK-P			PUD			PUD-N			PUD-W			GSD		
CL	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*
1	235	2.63	1	730	3.13	1	384	3.10	1	346	3.17	1	3783	3.31
2	251	2.01	2	856	2.56	2	457	2.50	2	399	2.62	2	4273	2.77
3	188	1.75	3	544	2.47	3	279	2.31	3	265	2.63	3	2985	2.51
4	128	1.67	4	343	2.51	4	172	2.34	4	171	2.69	4	1462	2.38
5	54	1.79	5	178	2.62	5	94	2.44	5	84	2.81	5	622	2.22
6.29	21	1.60	6	82	2.39	6	37	2.38	6	45	2.40	6	171	2.10
			7	25	2.45	7.36	14	2.31	7.32	22	2.60	7	45	1.88
			8.09	11	2.57							8.14	14	1.51
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.63	2.39	*a*	3.13	2.90	*a*	3.10	2.84	*a*	3.17	2.99	*a*	3.31	3.49
b	-0.30	-0.51	*b*	-0.14	-0.30	*b*	-0.18	-0.36	*b*	-0.13	-0.22	*b*	-0.28	-0.10
c		-0.09	*c*		-0.06	*c*		-0.08	*c*		-0.04	*c*		0.07
R^2	0.8990	0.9670	R^2	0.4544	0.8461	R^2	0.6469	0.9168	R^2	0.4703	0.6313	R^2	0.9279	0.9716



HK-P PUD GSD

PUD-N PUD-W

Figure 26. MAL applied to the triplet of the clause, phrase and word – inclusion of clausal heads governing at least one phrase.

When evaluating the impact of the inclusive approach, firstly, the proportion of the excluded words decreased to less than $4\,\%$ in the samples (see Table 26). Secondly, the standard of $R^2 \geq 0.90$ is reached in GSD for both the models, and in HK-P and PUD-N for M2. The fit of M1 in HK-P is only slightly below the standard, i.e. $R^2 = 0.8990$. As for PUD and PUD-W, even though the standard of $R^2 \geq 0.90$ is not met, the parameter b of M1 has negative values contrary to the previous results. We are fully aware that this inclusive approach can be regarded as trivial – MPL_1s increase by 1.0, while the increase in other MPL lowers with the increasing CL. The approach also suffers from other methodological drawbacks. For example, clausal heads with zero phrases remain excluded from the analysis. However, the approach, first and foremost, illustrates the serious impact the original approach to the clausal phrase has on $MPLs$.

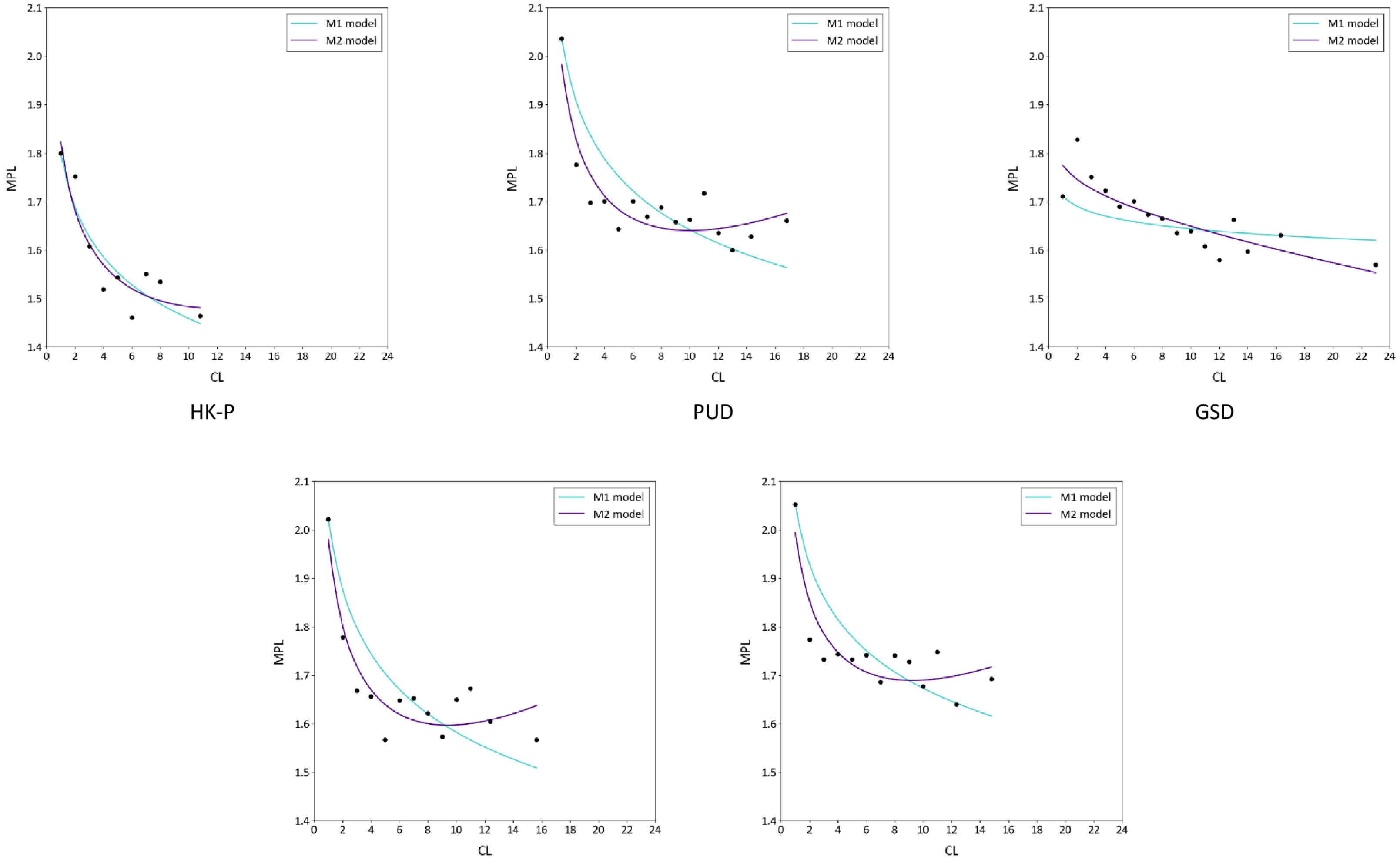
4.2.3 The linear dependency segment and word as constituents

Hypothesis: the longer the clause length measured in the number of linear dependency segments (LDS), the shorter the mean length of LDSs measured in words.

Finally, we present the results for the triplet including LDS as the constituent of the clause (Table 28 and Figure 27). CL stands for the clause length measured the in the number of LDS, $f(CL)$ for its frequency and MPL for the mean LDS length measured in the number of words. The parameters (a, b, c) and the coefficient of determination R^2 of both the models – the truncated model $y(x) = ax^b$ with the label M1 and the complete model $y(x) = ax^b e^{cx}$ with the label M2 – are shown in the table. In the case of M1, the mean LDS length of one-LDS clauses, i.e. MPL_1, is used the parameter a. Finally, if a value of MPL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 28. MAL applied to the triplet of the clause, linear dependency segment and word.

HK-P			PUD			PUD-N			PUD-W			GSD		
CL	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*
1	365	1.80	1	764	2.04	1	406	2.02	1	358	2.05	1	5044	1.71
2	222	1.75	2	637	1.78	2	354	1.78	2	283	1.77	2	3048	1.83
3	187	1.61	3	456	1.70	3	244	1.67	3	212	1.73	3	2502	1.75
4	103	1.52	4	295	1.70	4	144	1.66	4	151	1.74	4	1732	1.72
5	74	1.54	5	226	1.64	5	122	1.57	5	104	1.73	5	1255	1.69
6	38	1.46	6	162	1.70	6	71	1.65	6	91	1.74	6	833	1.70
7	14	1.55	7	116	1.67	7	60	1.65	7	56	1.69	7	583	1.67
8	11	1.53	8	97	1.69	8	43	1.62	8	54	1.74	8	339	1.67
10.80	10	1.46	9	69	1.66	9	31	1.57	9	38	1.73	9	216	1.64
			10	48	1.66	10	26	1.65	10	22	1.68	10	124	1.64
			11	37	1.72	11	15	1.67	11	22	1.75	11	91	1.61
			12	27	1.64	12.40	20	1.60	12.32	22	1.64	12	49	1.58
			13	15	1.60	15.64	14	1.57	14.79	19	1.69	13	26	1.66
			14.32	22	1.63							14	22	1.60
			16.82	11	1.66							16.33	18	1.63
												23	11	1.57
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.80	1.80	*a*	2.04	1.96	*a*	2.02	1.95	*a*	2.05	1.97	*a*	1.71	1.78
b	-0.09	-0.13	*b*	-0.09	-0.13	*b*	-0.11	-0.16	*b*	-0.09	-0.12	*b*	-0.02	-0.02
c		-0.01	*c*		-0.01	*c*		-0.02	*c*		-0.01	*c*		0.003
R^2	0.8447	0.8662	R^2	0.4706	0.8357	R^2	0.5549	0.8442	R^2	0.4340	0.7842	R^2	0.4249	0.7182



HK-P

PUD

GSD

PUD-N PUD-W

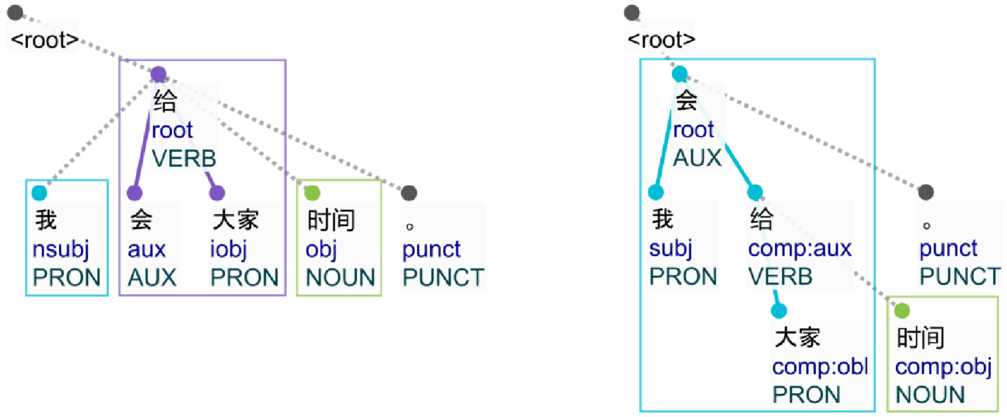
Figure 27. MAL applied to the triplet of the clause, linear dependency segment and word.

When taking the standard of $R^2 \geq 0.90$ into account, the hypothesis is rejected. Nevertheless, the parameter b of M1 has negative values, hence, R^2 reflects the decreasing trend in $MPLs$ (although in GSD is only indicated). HK-P shows the best fitting results for both the models ($R^2 = 0.8447$ for M1 and $R^2 = 0.8662$ for M2). As for the PUD samples, R^2 is exceeding the value of 0.80 (PUD and PUD-N) or approximating it (PUD-W) only when M2 is applied. GSD achieves the worst fitting results for both models. The second regime occurs again to a greater degree in all the samples (highlighted in yellow in Table 28).

To measure the clause in LDSs raises again the issue of the construct lengths being above the upper threshold of the short-term memory span (7 ± 2, Miller, 1956). CLs have, for example, up to 16.82 LDSs in PUD and up to 23 LDSs in GSD. The division of the clause into LDSs is governed not only by the dependency syntactic criterion but also by the linear neighbourhood. Hence, the clauses are more fragmented, or in other words, have higher numbers of LDSs compared to clausal phrases.[128] As a result, the mean LDS lengths reach slightly lower values than the mean lengths of the clausal phrases, even though LDSs include all clausal heads being previously excluded.

In addition, Mačutek, Čech and Courtin (2021) originally tested LDSs on a language material of the Surface-syntax Universal Dependencies (SUD, Gerdes et al., 2018), where the dependency relations are not necessarily governed by content words contrary to UD. Consequently, the inverted direction deepens syntactic structures, which results in lower numbers of LDSs in clauses. To illustrate the difference between UD and SUD, we use the same sentence (corresponding to one clause) annotated in both the frameworks and decompose it into LDSs (see Figure 28). The promotion of the content word 给 (*gěi*, ‘give’) in the UD version (on the left) flattens the syntactic structure and results in three LDSs while assigning the role of the root to the auxiliary verb 会 (*huì*, ‘be able to; be good at; be likely’) in SUD (on the right) deepens the structure and results in two LDSs.

[128] When comparing the sum of the clausal phrases and LDSs in each sample, HK-P contains $2{,}215$ clausal phrases and $2{,}674$ LDSs, PUD $7{,}092$ clausal phrases and $10{,}803$ LDSs, and GSD $31{,}697$ clausal phrases and $49{,}606$ LDSs.



我会给大家时间。

Wǒ huì gěi dàjiā shíjiān.

'I will give everyone time.'

Source: CoNLL-U Viewer (CoNLL-U Viewer), adjusted by the author

Figure 28. The example of decomposition of a sentence (sentence ID 677, HK-P treebank) annotated in UD (left) and SUD (right) into linear dependency segments.

The difference in the governance of the dependency relations raises the question of whether the LDS approach originally suggested for SUD syntactic structures contributes to or causes the wide scale of the clause lengths when applied to UD. Since SUD does not directly annotate the clausal dependency relations (see Chapter 4.1.4), we test both the annotations only on one-clause sentences divided into two samples. The first contains one-clause sentences from HK-P representing the spoken text type and the second includes one-clause sentences from PUD and GSD representing the written text type. We merged the sentences from PUD and GSD into one sample due to their low frequency. The results are presented in Table 29.

Table 29. MAL applied to the triplet of the clause, linear dependency segment and word – one-clause sentences in the UD and SUD frameworks.

	HK-P UD			HK-P SUD			PUD & GSD UD			PUD & GSD SUD	
CL	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*	*CL*	*f(CL)*	*MPL*
1	12	2.17	1	13	2.31	1.80	15	2.43	1.71	21	2.90
2	13	1.88	2	20	2.20	3	34	2.17	3	48	2.44
3	17	1.69	3	21	1.83	4	64	2.02	4	78	2.16
4	15	1.57	4.27	11	1.68	5	96	1.91	5	96	1.98
6.28	18	1.47	6.80	10	1.62	6	124	1.81	6	105	1.84
						7	82	1.76	7	78	1.84
						8	55	1.77	8	55	1.77
						9	40	1.68	9	30	1.76
						10	21	1.70	10	31	1.76
						11.31	29	1.70	11	11	1.65
						13.42	12	1.62	12.31	16	1.63
						17.10	10	1.71	16.08	13	1.76
	M1	M2		M1	M2		M1	M2		M1	M2
a	2.17	2.16	*a*	2.31	2.37	*a*	2.63*	2.86	*a*	3.25*	3.57
b	-0.22	-0.24	*b*	-0.19	-0.18	*b*	-0.19	-0.33	*b*	-0.28	-0.48
c		-0.01	*c*		0.01	*c*		-0.02	*c*		-0.04
R^2	0.9925	0.9935	R^2	0.8981	0.9087	R^2	0.9190	0.9861	R^2	0.9232	0.9859

*calculated by means of the NLREG software because CL_1 is pooled with CL_2 due to its insufficient frequency; as a result, MPL_1 cannot be used as the parameter a in this case (see Chapter 4.1.2)

Apart from one case being slightly below the standard of $R^2 \geq 0.90$ (i.e. HK-P SUD fitted by M1), the law comes into force and the hypothesis is corroborated for both the annotations and the samples. Even though shorter CLs have a slightly higher frequency in SUD, the CL scales approximate each other. The upper threshold of the short-term memory span (7 ± 2, Miller, 1956) is exceeded in the sample of PUD and GSD without regard to the annotation. As for the constituent lengths, LDSs in UD are shorter on average than in SUD. Nevertheless, both the annotation frameworks yield similar results and they seem to differ from each other to a minimal extent. However, due to the limited size of the samples, the conclusions are only tentative.

4.2.4 Summary of triplets on the clause level

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 (see Table 30).

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 CL 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 (see Table 30), 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.

Table 30. The parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained on the clausal level.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	clause-word-character									
a	2.08	1.91	1.87	1.79	1.89	1.79	1.85	1.79	1.49	1.57
b	-0.14	-0.20	-0.02	-0.02	-0.03	-0.05	-0.02	-0.003	0.05	0.04
c		-0.03		-0.003		-0.01		0.00		0.001
R^2	NA	0.5862	NA	0.1936	NA	0.2602	NA	0.0076	0.2368	0.5188
	clause-word-character – punctuation approach									
a	2.06	2.00	2.15	2.12	2.20	2.13	2.07	2.10	2.29	2.24
b	-0.12	-0.21	-0.08	-0.15	-0.09	-0.18	-0.06	-0.12	-0.11	-0.18
c		-0.02		-0.01		-0.02		-0.01		-0.01
R^2	0.4084	0.7812	0.1903	0.8047	NA	0.8715	0.4299	0.5473	0.3899	0.8593
	clause-phrase-word									
a	1.63	1.55	2.13	2.05	2.10	2.00	2.17	2.14	2.31	2.54
b	-0.07	-0.18	0.04	-0.01	0.01	-0.06	0.06	0.11	-0.12	0.19
c		-0.05		-0.02		-0.03		0.01		0.11
R^2	0.2555	0.4628	0.5078	0.6592	0.1598	0.5448	0.4049	0.4260	0.6044	0.9385
	clause-phrase-word – inclusion of clausal heads with at least one phrase									
a	2.63	2.39	3.13	2.90	3.10	2.84	3.17	2.99	3.31	3.49
b	-0.30	-0.51	-0.14	-0.30	-0.18	-0.36	-0.13	-0.22	-0.28	-0.10
c		-0.09		-0.06		-0.08		-0.04		0.07
R^2	0.8990	0.9670	0.4544	0.8461	0.6469	0.9168	0.4703	0.6313	0.9279	0.9716
	clause-linear dependency segment-word									
a	1.80	1.80	2.04	1.96	2.02	1.95	2.05	1.97	1.71	1.78
b	-0.09	-0.13	-0.09	-0.13	-0.11	-0.16	-0.09	-0.12	-0.02	-0.02
c		-0.01		-0.01		-0.02		-0.01		0.003
R^2	0.8447	0.8662	0.4706	0.8357	0.5549	0.8442	0.4340	0.7842	0.4249	0.7182

4.3 The phrase as the construct

4.3.1 The word and character as constituents

Hypothesis: the longer the length of a phrasal unit[129] measured in the number of words, the shorter the mean length of the words measured in (Chinese) characters[130].

4.3.1.1 The sentential phrase

Table 31 and Figure 29 present the results obtained when testing the sentential phrase (i.e. a phrasal subtree directly dependent on a root of a sentence) in the position of the construct. PL labels the phrase length measured the in the number of words, $f(PL)$ its frequency and MWL the mean word length measured in the number of (Chinese) characters. The lengths are fitted by the truncated model $y(x) = ax^b$ with the label M1 and by the complete model $y(x) = ax^b e^{cx}$ with the label M2. Their parameters (a, b, c) and the coefficient of determination R^2 are presented in the table. In case of M1, we use MWL_1, i.e. the mean word length of one-word phrases, as the parameter a.

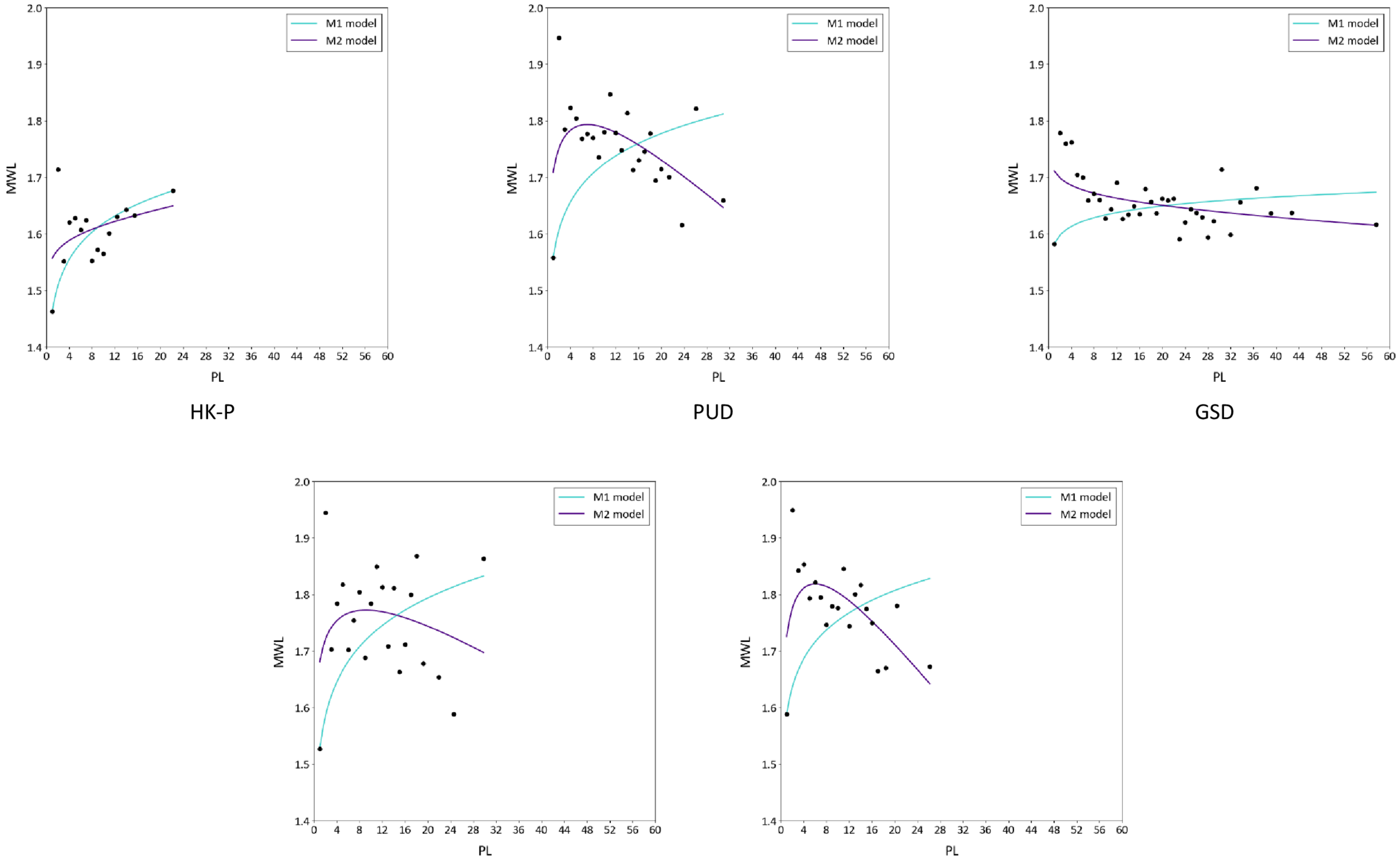
[129] I.e. sentential phrase, clausal phrase and linear dependency segment.

[130] Or syllables due to their close correspondence in Chinese (except for erization).

Table 31. MAL applied to the triplet of the sentential phrase, word and character.

HK-P			PUD			PUD-N			PUD-W			GSD		
PL	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*
1	775	1.46	1	1756	1.56	1	876	1.53	1	880	1.59	1	5649	1.58
2	131	1.71	2	386	1.95	2	180	1.94	2	206	1.95	2	2347	1.78
3	90	1.55	3	315	1.79	3	129	1.70	3	186	1.84	3	1641	1.76
4	60	1.62	4	267	1.82	4	117	1.78	4	150	1.85	4	1460	1.76
5	42	1.63	5	231	1.80	5	104	1.82	5	127	1.79	5	1116	1.70
6	48	1.61	6	210	1.77	6	94	1.70	6	116	1.82	6	943	1.70
7	38	1.62	7	168	1.78	7	74	1.75	7	94	1.79	7	733	1.66
8	26	1.55	8	137	1.77	8	55	1.80	8	82	1.75	8	590	1.67
9	20	1.57	9	109	1.74	9	52	1.69	9	57	1.78	9	450	1.66
10	20	1.57	10	111	1.78	10	51	1.78	10	60	1.78	10	416	1.63
11	18	1.60	11	75	1.85	11	35	1.85	11	40	1.85	11	306	1.64
12.40	20	1.63	12	58	1.78	12	29	1.81	12	29	1.74	12	251	1.69
14	12	1.64	13	51	1.75	13	29	1.71	13	22	1.80	13	226	1.63
15.46	13	1.63	14	41	1.81	14	25	1.81	14	16	1.82	14	190	1.63
22.18	11	1.68	15	36	1.71	15	20	1.66	15	16	1.78	15	141	1.65
			16	26	1.73	16	13	1.71	16	13	1.75	16	117	1.64
			17	25	1.75	17	15	1.80	17	10	1.66	17	102	1.68
			18	19	1.78	18	11	1.87	18.43	14	1.67	18	86	1.66
			19	15	1.69	19.18	11	1.68	20.38	13	1.78	19	62	1.64
			20	10	1.72	21.91	11	1.65	26.17	12	1.67	20	59	1.66

			21.33	15	1.70	24.55	11	1.59				21	46	1.66
			23.64	11	1.62	29.80	10	1.86				22	28	1.66
			26.08	13	1.82							23	34	1.59
			30.90	10	1.66							24	29	1.62
												25	21	1.64
												26	19	1.64
												27	20	1.63
												28	14	1.59
												29	16	1.62
												30.43	21	1.71
												32	12	1.60
												33.71	17	1.66
												36.55	11	1.68
												39.08	13	1.64
												42.70	10	1.64
												57.60	10	1.62
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.46	1.56	*a*	1.56	1.72	*a*	1.53	1.69	*a*	1.59	1.74	*a*	1.58	1.71
b	0.04	0.01	*b*	0.04	0.04	*b*	0.05	0.04	*b*	0.04	0.05	*b*	0.01	-0.01
c		-0.001	*c*		0.01	*c*		0.004	*c*		0.01	*c*		0.0003
R^2	NA	0.1694	R^2	NA	0.2422	R^2	NA	0.0657	R^2	NA	0.3265	R^2	NA	0.2103



HK-P

PUD

GSD

PUD-N PUD-W

Figure 29. MAL applied to the triplet of the sentential phrase, word and character.

Firstly, the values of the coefficient of determination R^2 are either not available (for M1) or unsatisfactory (for M2). Hence, this unit triplet does not corroborate the hypothesis. On the contrary, the hypothesis was not rejected when the sentential phrase was the direct constituent of the sentence. These two opposite results amplify the need to test a respective unit in all possible positions in the unit hierarchy.

Secondly, the triplet suffers from PLs being above the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) in all the samples. Determining the phrase as a whole subtree directly dependent on a root leads to excessive lengths. For example, PLs have up to 22.18 words in HK-P, 30.90 in PUD and 57.60 in GSD. When we applied the law to the triplet of the sentence, phrase and word, the mean phrase lengths belonging to one-phrase sentences, i.e. MPL_1, also crossed this upper threshold (except for HK-P). Both the results of the sentential phrase being first the direct constituent of the sentence and then the construct itself 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).

Finally, the word being the constituent shows its lengths fluctuating in a narrow range between one and two (Chinese) characters. As discussed in Chapter 4.2.1, these word lengths are the most frequent in our samples and Chinese. Hence, their prevalence might limit the law to manifest itself. In addition, involving the word length raises the issue of the inclusion of non-Chinese graphemes.[131] However, we decided not to test the impact of their exclusion on the results due to the drawbacks of this triplet mentioned above.

[131] We remind the reader that one grapheme in a Chinese word, i.e. Chinese character, roughly corresponds to one syllable, whereas one grapheme in a non-Chinese word usually represents a letter, numeral, or symbol.

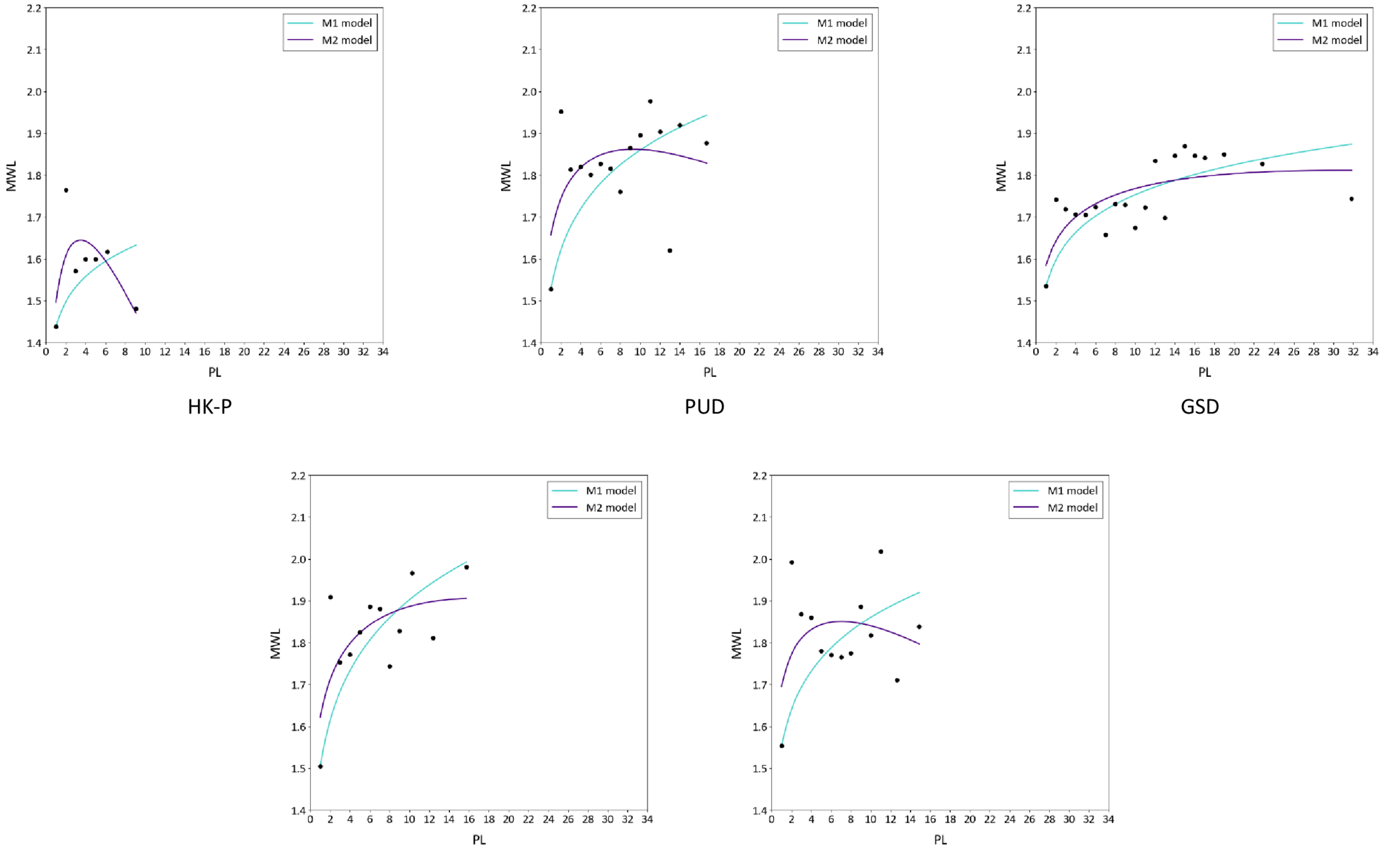
4.3.1.2 The clausal phrase

The results yielded by testing the clausal phrase are presented in Table 32 and Figure 30. PL denotes the phrase length measured in the number of words, $f(PL)$ its frequency and MWL the mean word length measured in the number of (Chinese) characters. Both the models are applied – the truncated model $y(x) = ax^b$ labelled as M1 and the complete model $y(x) = ax^b e^{cx}$ labelled as M2. Their parameters (a, b, c) and the coefficient of determination R^2 can be found in the table. We again use MWL_1 as the parameter a when fitting M1 to the data.

Table 32. MAL applied to the triplet of the clausal phrase, word and character.

HK-P			PUD			PUD-N			PUD-W			GSD		
PL	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*
1	1675	1.44	1	4418	1.53	1	2358	1.50	1	2060	1.55	1	17972	1.53
2	263	1.76	2	839	1.95	2	409	1.91	2	430	1.99	2	5464	1.74
3	157	1.57	3	605	1.81	3	288	1.75	3	317	1.87	3	2918	1.72
4	55	1.60	4	378	1.82	4	174	1.77	4	204	1.86	4	1910	1.71
5	32	1.60	5	275	1.80	5	131	1.82	5	144	1.78	5	1228	1.71
6.18	22	1.62	6	176	1.83	6	85	1.89	6	91	1.77	6	831	1.72
9.09	11	1.48	7	123	1.82	7	53	1.88	7	70	1.77	7	506	1.66
			8	84	1.76	8	39	1.74	8	45	1.78	8	264	1.73
			9	56	1.87	9	20	1.83	9	36	1.89	9	184	1.73
			10	48	1.90	10.26	27	1.97	10	28	1.82	10	128	1.67
			11	27	1.98	12.36	22	1.81	11	20	2.02	11	91	1.72
			12	19	1.90	15.75	12	1.98	12.62	13	1.71	12	61	1.83
			13	16	1.62				14.88	16	1.84	13	40	1.70
			14	15	1.92							14	20	1.85
			16.69	13	1.88							15	23	1.87
												16	11	1.85
												17	10	1.84
												18.93	15	1.85
												22.80	10	1.83
												31.82	11	1.74

	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.44	1.57	*a*	1.53	1.67	*a*	1.50	1.63	*a*	1.55	1.72	*a*	1.53	1.59
b	0.06	0.18	*b*	0.09	0.09	*b*	0.10	0.08	*b*	0.08	0.08	*b*	0.06	0.05
c		0.05	*c*		0.01	*c*		0.005	*c*		0.01	*c*		0.002
R^2	NA	0.4521	R^2	NA	0.2340	R^2	0.1894	0.4653	R^2	NA	0.1413	R^2	0.3810	0.5070



HK-P PUD GSD

PUD-N PUD-W

Figure 30. MAL applied to the triplet of the clausal phrase, word and character.

The phrase being the constituent to the clause (not to the sentence) does not unambiguously improve the results. Values of the coefficient of determination R^2 are either not available (for M1) or highly unsatisfactory (for M1 and M2). The data points of $MWLs$ are scattered and the fitting curves are rather rising.

The inclusion of the clause into the hierarchy of language units reduced the scale of PLs compared to the sentential phrase. Nonetheless, the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) remains exceeded except for HK-P. The clausal phrases, for example, have lengths up to 16.69 words in PUD and 31.82 in GSD. As for the constituent, the narrow range of one to two (Chinese) characters in which $MWLs$ fluctuate does not change. Hence, the word length distribution in Chinese still appears to limit the law.

As mentioned in the previous triplet, when the word measured in Chinese characters becomes the constituent, the issue of words fully or partly consisting of non-Chinese graphemes arises. Even though these words represent only $3.60\ \%$ of all word tokens in PUD and $3.03\ \%$ in GSD, we exclude all phrases which contain at least one non-Chinese grapheme and rerun the analysis. The results of both the approaches – inclusive and exclusive – are presented in Table 33. As can be seen, the exclusion does not considerably influence the results, similarly to the clausal level (Chapter 4.2.1). The goodness-of-fit remains highly unsatisfactory.

Table 33. The number of phrases $n(P)$, the parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained by the inclusion and exclusion of phrases containing non-Chinese graphemes.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	inclusion of phrases containing non-Chinese graphemes									
n(P)	2215		7092		3618		3474		31697	
a	1.44	1.57	1.53	1.67	1.50	1.63	1.55	1.72	1.53	1.59
b	0.06	0.18	0.09	0.09	0.10	0.08	0.08	0.08	0.06	0.05
c		0.05		0.01		0.005		0.01		0.002
R^2	NA	0.4521	NA	0.2340	0.1894	0.4653	NA	0.1413	0.3810	0.5070
	exclusion of phrases containing non-Chinese graphemes									
n(P)	2212		6627		3376		3251		29041	
a	1.44	1.57	1.50	1.62	1.46	1.56	1.54	1.68	1.48	1.55
b	0.06	0.17	0.06	0.10	0.05	0.08	0.06	0.09	0.04	0.02
c		0.05		0.02		0.01		0.02		- 0.0003
R^2	NA	0.4383	NA	0.3160	NA	0.1921	NA	0.3051	0.1762	0.4762

Another issue raises in connection with the phrase lengths and their frequencies $f(PL)$. As addressed in Chapter 1.4, unit tokens reflect the competition between the Brevity law and the Menzerath-Altmann law. According to the former law, unit lengths and their frequencies are negatively correlated, i.e. shorter units are more frequently used. Suppose the frequency of the unit usage is involved. In that case, the Brevity law may affect not only the construct but also the constituent and impose double constraints on the Menzerath-Altmann law. One-word phrases account for 75.62 % of all phrases in HK-P, 62.30 % in PUD and 56.70 % in GSD, and their mean word lengths have the lowest values contradicting the Menzerath-Altmann law. For example, HK-P contains 1675 one-word phrase tokens, 981 of which consist of a word having only one character, e.g. 我 (*wǒ*, 'I, me') occurring 144 times in this sample. If we consider only types, the number of these one-word phrases consisting of one character is reduced from 981 to 96 (i.e. one-word phrase 我, *wǒ*, 'I, me', is counted only once). The frequency of usage appears to lower mean word lengths considerably. Moreover, phrases having two and more words might be similarly affected. Due to a high probability that the results are biased towards phrase tokens, we decided to apply the Menzerath-Altmann law to phrase types. Table 34 includes the results. Although the goodness-of-fit does not reach the standard of $R^2 \geq 0.90$ and the hypothesis is rejected, testing only the phrase types considerably influences the results. The parameter b of M1 has negative values in all the samples, contrary to positive values yielded by the phrase tokens.

Table 34. The number of phrases $n(P)$, the parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained by testing phrase tokens and types.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	phrase tokens									
n(P)	2215		7092		3618		3474		31697	
a	1.44	1.57	1.53	1.67	1.50	1.63	1.55	1.72	1.53	1.59
b	0.06	0.18	0.09	0.09	0.10	0.08	0.08	0.08	0.06	0.05
c		0.050		0.01		0.00		0.01		0.00
R²	NA	0.4521	NA	0.2340	0.1894	0.4653	NA	0.1413	0.3810	0.5070
	phrase types									
n(P)	745		3715		1901		2033		16629	
a	1.78	1.78	2.23	2.15	2.13	2.05	2.16	2.12	2.26	2.03
b	-0.07	-0.07	-0.09	-0.16	-0.07	-0.15	-0.08	-0.15	-0.10	-0.12
c		0.001		-0.02		-0.02		-0.02		-0.01
R²	0.8450	0.8454	0.1279	0.5733	NA	0.6250	0.3862	0.6311	NA	0.4009

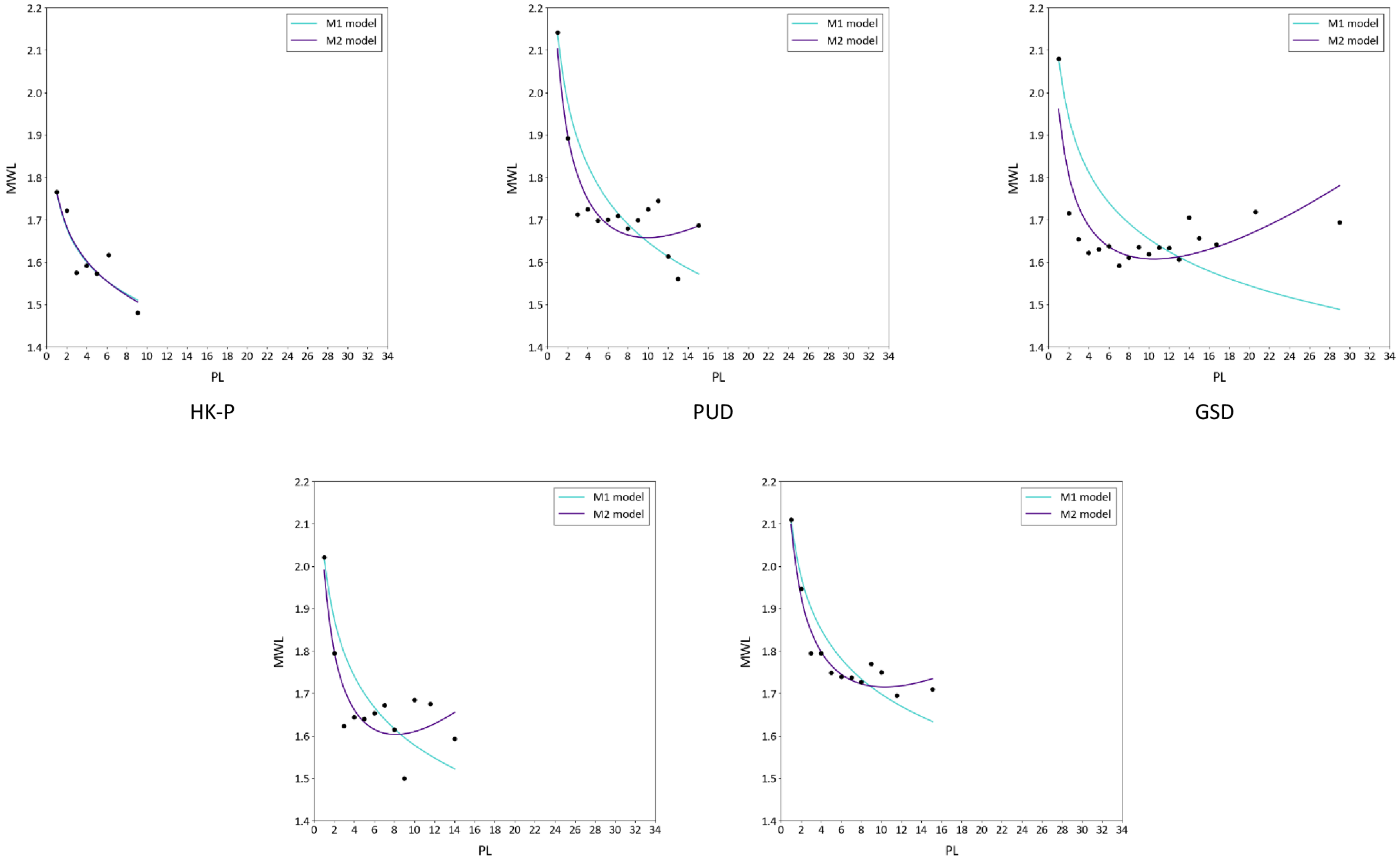
Due to the fact that the proportion of phrases including at least one non-Chinese grapheme increases with the types, we also apply the law to phrase types consisting only of Chinese characters to test whether the exclusion has a higher impact than on the previous results.[132] Table 35 and Figure 31 present the results.

[132] In the case of word tokens and word types, the proportion of words consisting of at least one non-Chinese grapheme increases from 3.60 % to 8.86 % in PUD and from 3.03 % to 9.24 % in GSD.

Table 35. MAL applied to the triplet of the clausal phrase, word and character – Chinese phrase types.

HK-P			PUD			PUD-N			PUD-W			GSD		
PL	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*
1	311	1.77	1	1105	2.14	1	656	2.02	1	650	2.11	1	3795	2.08
2	189	1.72	2	662	1.89	2	322	1.80	2	357	1.95	2	3974	1.72
3	125	1.58	3	523	1.71	3	254	1.62	3	270	1.80	3	2428	1.65
4	54	1.59	4	333	1.73	4	153	1.64	4	180	1.79	4	1532	1.62
5	30	1.57	5	218	1.70	5	101	1.64	5	117	1.75	5	984	1.63
6.18	22	1.62	6	135	1.70	6	62	1.65	6	73	1.74	6	568	1.64
9.09	11	1.48	7	95	1.71	7	41	1.67	7	54	1.74	7	351	1.59
			8	57	1.68	8	24	1.61	8	33	1.73	8	207	1.61
			9	38	1.70	9	10	1.50	9	28	1.77	9	135	1.64
			10	35	1.73	10	13	1.68	10	22	1.75	10	99	1.62
			11	20	1.75	11.57	14	1.68	11.52	21	1.69	11	68	1.64
			12	11	1.61	14	12	1.59	15.10	10	1.71	12	45	1.63
			13	10	1.56							13	26	1.61
			15.06	16	1.69							14	16	1.71
												15	13	1.66
												16.73	15	1.64
												20.60	10	1.72
												29	11	1.69

	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.77	1.77	*a*	2.14	2.07	*a*	2.02	1.95	*a*	2.11	2.07	*a*	2.08	1.94
b	-0.07	-0.07	*b*	-0.11	-0.17	*b*	-0.11	-0.18	*b*	-0.09	-0.14	*b*	-0.10	-0.14
c		0.001	*c*		-0.02	*c*		-0.02	*c*		-0.01	*c*		-0.01
R^2	0.8181	0.8191	R^2	0.6174	0.8367	R^2	0.4939	0.7898	R^2	0.7833	0.9439	R^2	NA	0.7283



HK-P PUD GSD

PUD-N PUD-W

Figure 31. MAL applied to the triplet of the clausal phrase, word and character – Chinese phrase types.

Values of the coefficient of determination R^2 obtained by fitting M1 and M2 to the modified data considerably increased except for HK-P fitted by both the models and GSD by M1. Despite these exceptions, testing phrase types consisting only of Chinese characters achieves better fitting results than testing all phrase types (see Table 34). Hence, the homogeneity of the samples appears to be another important factor when the types are analysed. Only HK-P shows the opposite trend – better results are achieved when analysing all phrase types. Its parameter a of M1, i.e. MWL_1, has a slightly higher value in this case. However, the value is biased by the inclusion of the expression 'Declaration_of_Renunciation_of_British_Citizenship', annotated as one word in UD and having the length of 50 graphemes. The homogeneity of the samples shows its importance again – its lower degree can lead to better results at the cost of biased data.

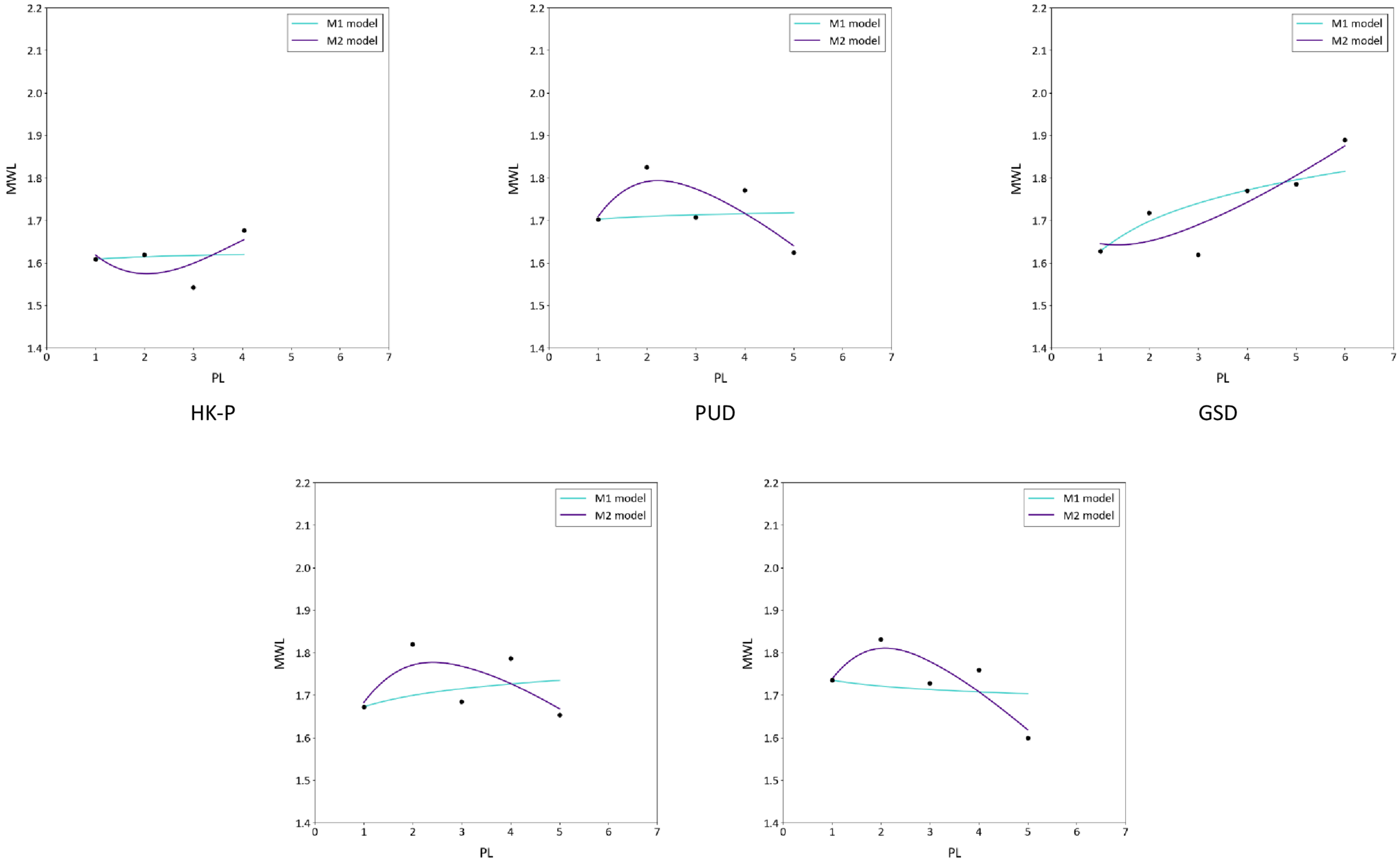
On the one hand, the goodness-of-fits do not meet the standard of $R^2 \geq 0.90$ and the hypothesis is rejected except for PUD-W fitted by M2. On the other hand, MWL_1s have the highest values following the Menzerath-Altmann law and the menzerathian decreasing trend appears compared to the initial results. Moreover, the overall worse fit might result from the drawbacks outlined above, i.e. the wide scale of construct lengths and the word length distribution in Chinese.

4.3.1.3 The linear dependency segment

The last results on this level come from testing the linear dependency segment (LDS) as the construct and can be found in Table 36 and Figure 32. PL is used for the LDS length measured in the number of words, $f(PL)$ for its frequency and MWL for the mean word length measured in the number of (Chinese) characters. We fit the data with both the models – M1 denoting the truncated model $y(x) = ax^b$ and M2 denoting the complete model $y(x) = ax^b e^{cx}$. Their parameters (a, b, c) and the coefficient of determination R^2 are included in the table. M1 model is fitted to the data while using MWL_1 as the parameter a. Finally, if a value of MWL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 36. MAL applied to the triplet of the linear dependency segment, word and character.

HK-P			PUD			PUD-N			PUD-W			GSD		
PL	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*
1	1393	1.61	1	5155	1.70	1	2673	1.67	1	2482	1.74	1	22813	1.63
2	953	1.62	2	3911	1.83	2	1920	1.82	2	1991	1.83	2	19780	1.72
3	300	1.54	3	1445	1.71	3	676	1.68	3	769	1.73	3	5610	1.62
4.04	28	1.68	4	259	1.77	4	109	1.79	4	150	1.76	4	1241	1.77
			5	33	1.62	5	15	1.65	5	18	1.60	5	150	1.79
												6	12	1.89
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.61	1.50	*a*	1.70	1.87	*a*	1.67	1.82	*a*	1.74	1.92	*a*	1.63	1.56
b	0.005	-0.15	*b*	0.01	0.20	*b*	0.02	0.18	*b*	-0.01	0.20	*b*	0.06	-0.07
c		-0.08	*c*		0.09	*c*		0.08	*c*		0.10	*c*		-0.05
R^2	0.0216	0.3666	R^2	NA	0.6185	R^2	NA	0.4079	R^2	0.0859	0.7826	R^2	0.6079	0.7892



HK-P PUD GSD

PUD-N PUD-W

Figure 32. MAL applied to the triplet of the linear dependency segment, word and character.

As demonstrated by fitting curves in Figure 32, the goodness-of-fit between the models and the data is unsatisfactory and the hypothesis is rejected for LDS. The data points of $MWLs$ from PUD and its versions show a zig-zag tendency (cf. Roukk, 2007) and the tendency of $MWLs$ in HK-P and GSD is even increasing.

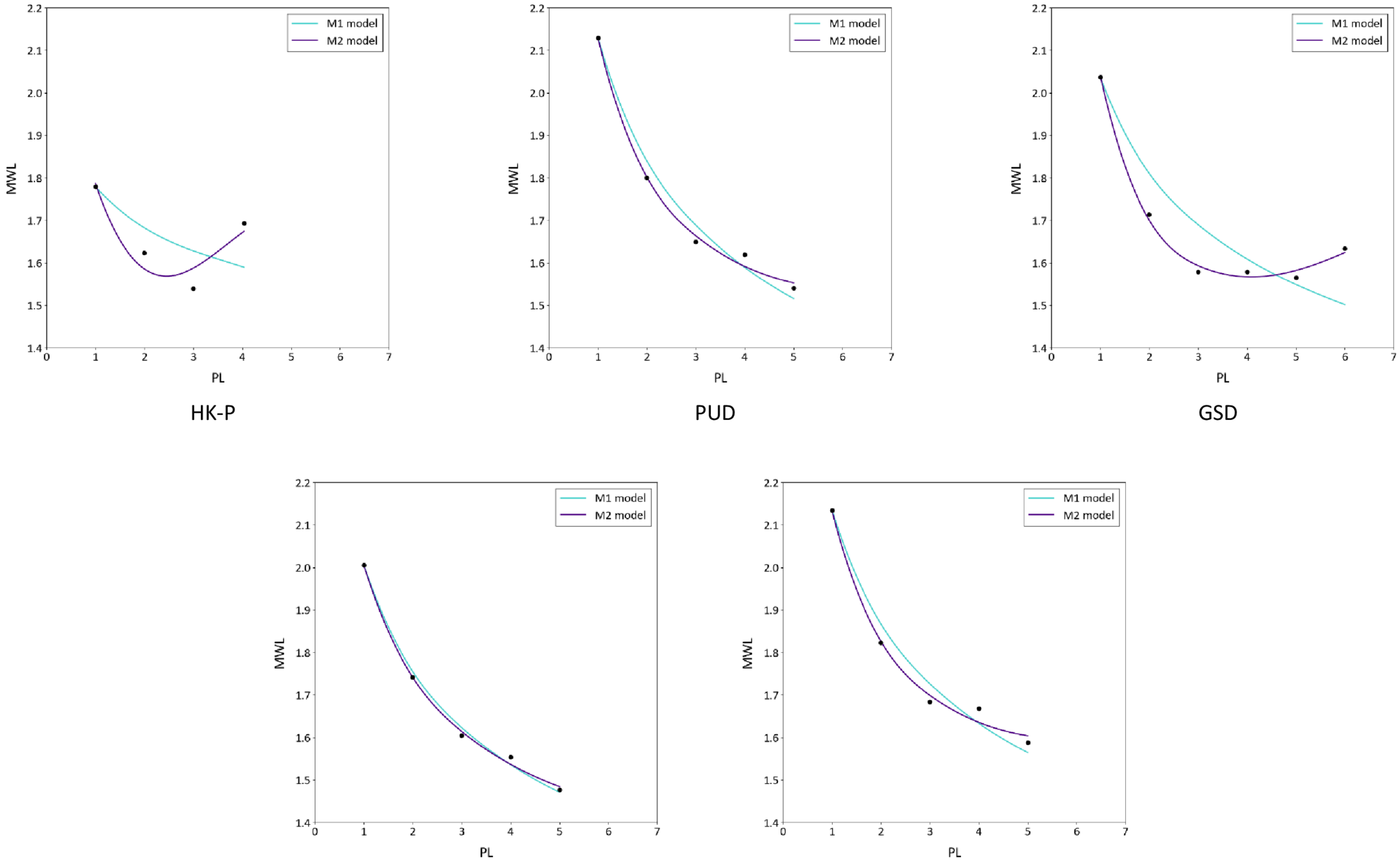
Despite the poor results, we finally achieve PLs which are in agreement with the upper threshold of the short-term memory span (i.e. 7 ± 2, Miller, 1956) in all the samples. The LDS lengths range from one to five words in HK-P (for un-pooled data), PUD and its versions and from one to six in GSD. These short lengths mainly result from the determination of LDS which takes the dependency syntactic criterion and the word order into account. As a result, clauses are fragmented to a higher degree than in the case of the clausal phrase, i.e. they consist of a higher number of LDSs shorter in words (see Chapter 4.1.4 and 4.2.3). As for the constituent, the word lengths are again between one and two (Chinese) characters on average.

The question arises whether the word length distribution in Chinese is the decisive factor or the law is limited by the frequency of the unit usage (i.e. by the Brevity law). Similarly to the clausal phrase, LDS faces the prevalence of one-word LDS tokens, which account for $52.09\ \%$ in HK-P, $47.72\ \%$ in PUD and $45.99\ \%$ in GSD. Their frequency and possible short lengths of their words might considerably decrease the mean word lengths and bias the results. For this reason, we follow the previous approach and apply the law to LDS types consisting only of Chinese characters. The obtained results are presented in Table 37 and Figure 33.[133]

[133] As demonstrated in the case of the clausal phrase, testing phrase types consisting only of Chinese characters achieved the best fitting results. Hence, we decided to follow this combined approach.

Table 37. MAL applied to the triplet of the linear dependency segment, word and character – types of Chinese linear dependency segment.

HK-P			PUD			PUD-N			PUD-W			GSD		
PL	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*	*PL*	*f(PL)*	*MWL*
1	362	1.78	1	1784	2.13	1	1017	2.01	1	1070	2.13	1	6018	2.04
2	654	1.62	2	2982	1.80	2	1516	1.74	2	1577	1.82	2	13494	1.71
3	271	1.54	3	1291	1.65	3	599	1.60	3	699	1.68	3	4806	1.58
4.04	27	1.69	4	208	1.62	4	89	1.55	4	119	1.67	4	993	1.58
			5	30	1.54	5	13	1.48	5	17	1.59	5	126	1.57
												6	10	1.63
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	1.78	1.51	*a*	2.13	2.04	*a*	2.01	1.97	*a*	2.13	2.04	*a*	2.04	1.84
b	-0.08	-0.42	*b*	-0.21	-0.30	*b*	-0.19	-0.23	*b*	-0.19	-0.29	*b*	-0.17	-0.41
c		-0.17	*c*		-0.04	*c*		-0.02	*c*		-0.04	*c*		-0.10
R^2	0.2935	0.8661	R^2	0.9790	0.9948	R^2	0.9949	0.9975	R^2	0.9699	0.9916	R^2	0.7563	0.9944



HK-P

PUD

GSD

PUD-N PUD-W

Figure 33. MAL applied to the triplet of the linear dependency segment, word and character – types of Chinese linear dependency segment.

First and foremost, the coefficient of determination R^2 meets the standard of $R^2 \geq 0.90$ in PUD and its versions (M1 and M2) and GSD (M2), and the hypothesis is corroborated in these cases. As regards HK-P, the fit of M1 is highly unsatisfactory and the fit of M2 reaches a considerably higher value, i.e. $R^2 = 0.8661$, because it includes the increase of the last MWL (see Figure 33). Nevertheless, the first three $MWLs$ clearly decrease and the last MWL represents a pooled value of a low frequency. GSD also shows a worse fit of M1, i.e. $R^2 = 0.7563$, but the last MWL_6, which as affected by the second regime, suffers from an extremely low frequency (it occurs only ten times in more than 25k LDSs and, in case of its omission, R^2 increases to 0.9076). Apart from the changes in the goodness-of-fit, MWL_1 being the parameter a of M1 reaches the highest values following the Menzerath-Altmann law and the parameter b of M1 has negative values in all the samples.

To sum it up, $MWLs$ decrease in the menzerathian trend despite the limited word length distribution in our samples and generally in Chinese, which initially appeared to be the boundary condition for the law. Hence, the decisive factor for the Menzerath-Altmann law to come into force is its application to the unit types. Moreover, the drawbacks that the clausal phrase has not overcome do not apply to LDS. Their lengths do not exceed the upper threshold of the short-term memory span.

4.3.2 Summary of triplets on the phrase level

The chapter presents results obtained by analysing three different units being the construct on the phrasal level. 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 $MWLs$ 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 (see Table 38).

Table 38. The parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained on the phrasal level.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	sentential phrase-word-character									
a	1.46	1.56	1.56	1.72	1.53	1.69	1.59	1.74	1.58	1.71
b	0.04	0.01	0.04	0.04	0.05	0.04	0.04	0.05	0.01	-0.01
c		-0.001		0.01		0.004		0.01		0.0003
R^2	NA	0.1694	NA	0.2422	NA	0.0657	NA	0.3265	NA	0.2103
	clausal phrase-word-character – all tokens									
a	1.44	1.57	1.53	1.67	1.50	1.63	1.55	1.72	1.53	1.59
b	0.06	0.18	0.09	0.09	0.10	0.08	0.08	0.08	0.06	0.05
c		0.05		0.01		0.005		0.01		0.002
R^2	NA	0.4521	NA	0.2340	0.1894	0.4653	NA	0.1413	0.3810	0.5070

clausal phrase-word-character – Chinese tokens									
1.44	1.57	1.50	1.62	1.46	1.56	1.54	1.68	1.48	1.55
0.06	0.17	0.06	0.10	0.05	0.08	0.06	0.09	0.04	0.02
	0.05		0.02		0.01		0.02		-0.0003
NA	0.4383	NA	0.3160	NA	0.1921	NA	0.3051	0.1762	0.4762
clausal phrase-word-character – all types									
1.78	1.78	2.23	2.15	2.13	2.05	2.16	2.12	2.26	2.03
-0.07	-0.07	-0.09	-0.16	-0.07	-0.15	-0.08	-0.15	-0.10	-0.12
	0.001		-0.02		-0.02		-0.02		-0.01
0.8450	0.8454	0.1279	0.5733	NA	0.6250	0.3862	0.6311	NA	0.4009
clausal phrase-word-character – Chinese types									
1.77	1.77	2.14	2.07	2.02	1.95	2.11	2.07	2.08	1.94
-0.07	-0.07	-0.11	-0.17	-0.11	-0.18	-0.09	-0.14	-0.10	-0.14
	0.001		-0.02		-0.02		-0.01		-0.01
0.8181	0.8191	0.6174	0.8367	0.4939	0.7898	0.7833	0.9439	NA	0.7283
linear dependency segment-word-character – all tokens									
1.61	1.50	1.70	1.87	1.67	1.82	1.74	1.92	1.63	1.56
0.005	-0.15	0.01	0.20	0.02	0.18	-0.01	0.20	0.06	-0.07
	-0.08		0.09		0.08		0.10		-0.05
0.0216	0.3666	NA	0.6185	NA	0.4079	0.0859	0.7826	0.6079	0.7892
linear dependency segment-word-character – Chinese types									
1.78	1.51	2.13	2.04	2.01	1.97	2.13	2.04	2.04	1.84
-0.08	-0.42	-0.21	-0.30	-0.19	-0.23	-0.19	-0.29	-0.17	-0.41
	-0.17		-0.04		-0.02		-0.04		-0.10
0.2935	0.8661	0.9790	0.9948	0.9949	0.9975	0.9699	0.9916	0.7563	0.9944

4.4 The word as the construct

4.4.1 The character and component as constituents

Hypothesis: the longer the word length[134] measured in the number of Chinese characters[135], the shorter the mean length of the characters measured in components.

The results of word tokens and word types are presented in Table 39 and Table 40 accordingly, and in Figure 34. WL labels the word length measured the in the number of Chinese characters, $f(WL)$ its frequency and MCL the mean character length measured in the number of components. We fit the lengths with M1 standing for the truncated model $y(x) = ax^b$ and with M2 standing for the complete model $y(x) = ax^b e^{cx}$. The parameters (a, b, c) and the coefficient of determination R^2 of both the models are shown in the tables. The parameter a of M1 equals a mean character length of one-character words, i.e. MCL_1. Finally, if a value of MCL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

[134] 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.

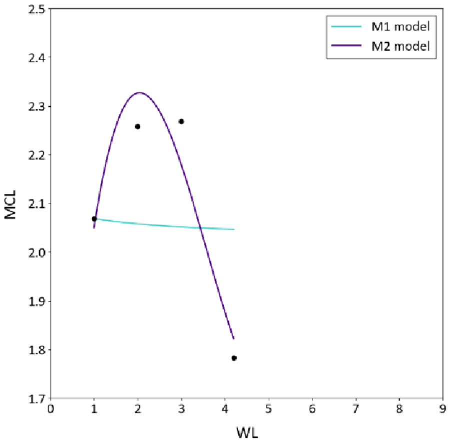
[135] We measure the word length in Chinese characters which correspond to syllables except for erization.

Table 39. MAL applied to the triplet of the word, character and component – tokens.

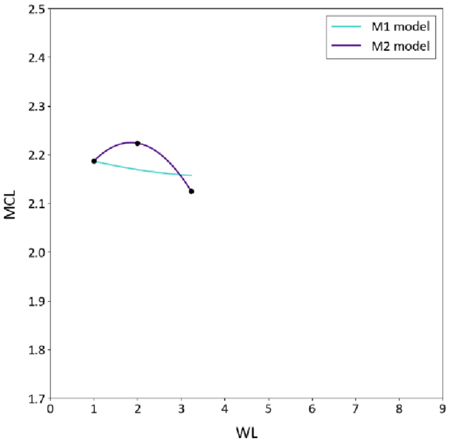
HK-P			PUD			PUD-N			PUD-W			GSD		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	1989	2.07	1	7784	1.87	1	3946	1.91	1	3838	1.83	1	36504	1.93
2	2155	2.26	2	8683	2.25	2	4195	2.26	2	4488	2.24	2	42058	2.24
3	149	2.27	3	1020	2.28	3	441	2.29	3	579	2.27	3	1538	2.29
4.20	10	1.78	4	274	2.30	4	94	2.34	4	180	2.28	4	707	2.17
			5	67	2.33	5.26	23	2.29	5	49	2.33	5	126	2.34
			6.25	16	2.24				6.27	11	2.28	6	30	2.41
												8	15	2.30
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.07	2.87	*a*	1.87	2.06	*a*	1.91	2.11	*a*	1.83	2.01	*a*	1.93	2.03
b	-0.01	0.67	*b*	0.14	0.34	*b*	0.14	0.35	*b*	0.15	0.33	*b*	0.11	0.19
c		0.34	*c*		0.09	*c*		0.10	*c*		0.08	*c*		0.03
R^2	NA	0.9055	R^2	0.4643	0.9567	R^2	0.6465	0.9749	R^2	0.5677	0.9414	R^2	0.5571	0.7748

Table 40. MAL applied to the triplet of the word, character and component – types.

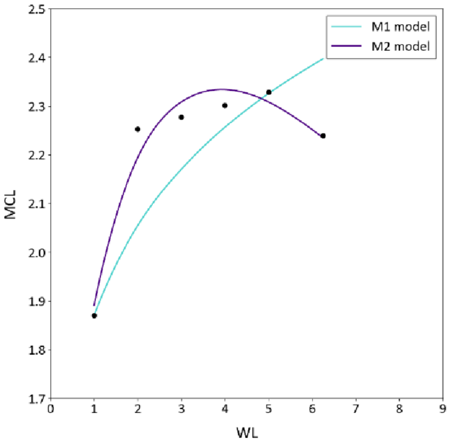
HK-P			PUD			PUD-N			PUD-W			GSD		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	198	2.19	1	469	2.26	1	339	2.22	1	332	2.20	1	1485	2.47
2	532	2.22	2	3456	2.31	2	2090	2.31	2	2138	2.29	2	12715	2.31
3.23	48	2.12	3	742	2.31	3	347	2.32	3	429	2.29	3	953	2.32
			4	207	2.31	4	77	2.35	4	135	2.29	4	509	2.27
			5	53	2.37	5.26	23	2.29	5	36	2.37	5	111	2.34
			6.25	16	2.24				6.27	11	2.28	6	27	2.43
												8	15	2.30
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.19	2.43	*a*	2.26	2.31	*a*	2.22	2.29	*a*	2.20	2.24	*a*	2.47	2.40
b	-0.01	0.18	*b*	0.01	0.07	*b*	0.03	0.11	*b*	0.03	0.07	*b*	-0.04	-0.08
c		0.11	*c*		0.02	*c*		0.03	*c*		0.02	*c*		-0.02
R^2	0.1964	>0.99	R^2	NA	0.4341	R^2	0.4632	0.9278	R^2	0.4946	0.6648	R^2	0.0039	0.3960



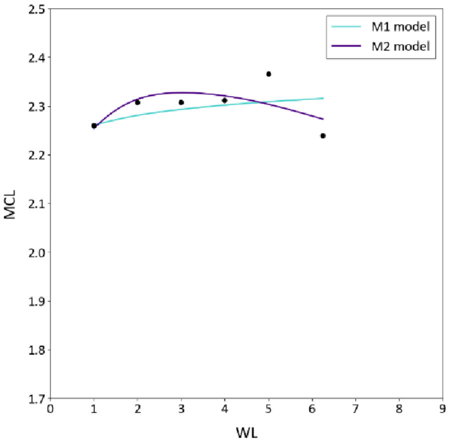
HK-P – tokens



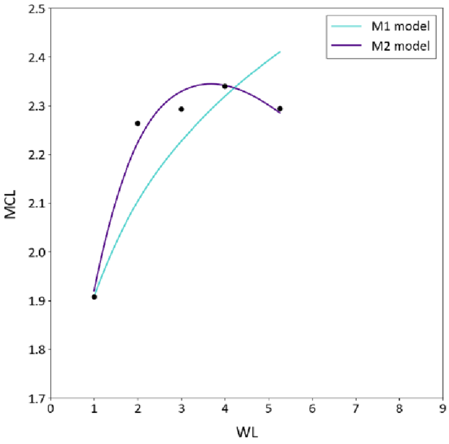
HK-P – types



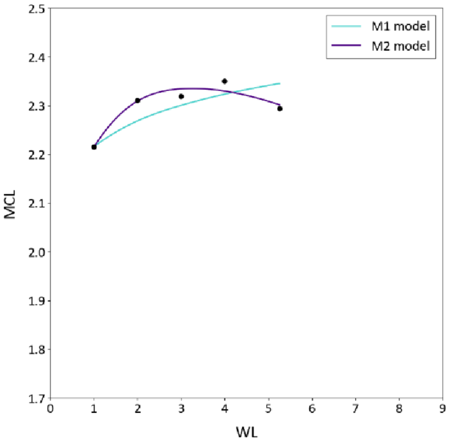
PUD – tokens



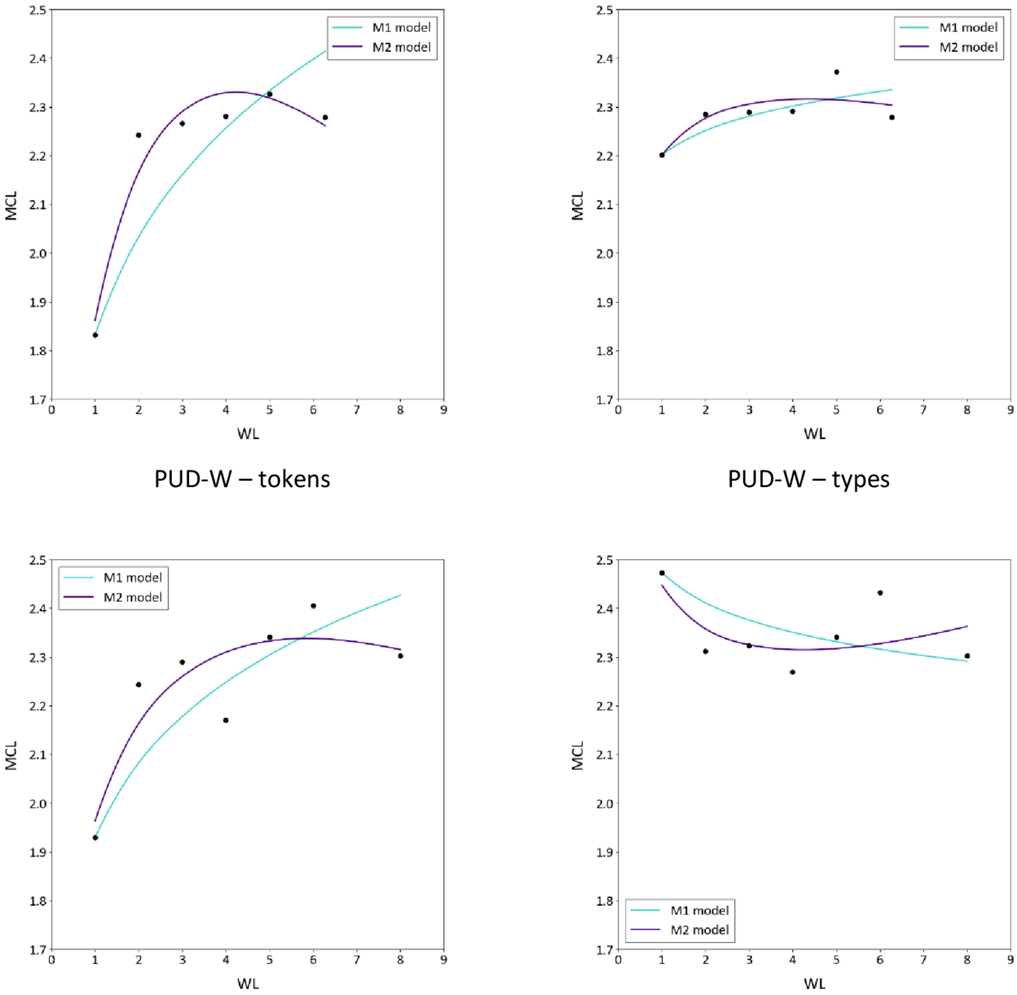
PUD – types



PUD – tokens



PUD-N – types



PUD-W – tokens

PUD-W – types

GSD – tokens

GSD – types

Figure 34. MAL applied to the triplet of the word, character and component – tokens and types.

To measure the word in Chinese characters and the Chinese character in components brings unsatisfactory results for both – the tokens and the types – and the hypothesis is rejected. In the case of the tokens, MCL_1s have a lower value than MCL_2 or even the lowest, and the rest of the mean character lengths rather increase while reaching their peak with next to the last $MCLs$. Hence, the fit of M1 is unsatisfactory (values of the parameter b are mostly positive), and M2 fits the increase in $MCLs$. The analysis of the types flattens the fitting curves but does not eliminate the increasing trend. In the case of the PUD samples, MCL_1s keep lower or the lowest values than MCL_2s and the peak of the fitting curves remains shifted towards the next to last $MCLs$. The only exception is GSD where MCL_1 is the highest and the next to last MCL approximates its value. Nevertheless, the fit of both the models is unsatisfactory in these samples. As for HK-P, we do not evaluate the results of the types. The sample has only three different word lengths after pooling the data, which are moreover fitted by three parameters in the case of M2 (i.e. $R^2 > 0.99$). As regards the short-term memory span, neither the scale of WL nor the scale of MCL exceeds its upper threshold (i.e. 7 ± 2, Miller, 1956).

The question arises as to what prevents the law from coming into force on the word level. Firstly, the news and/or Wikipedia articles in PUD and GSD contain foreign proper nouns (e.g. anthroponyms or toponyms), which might distort the sample homogeneity. The proper nouns are usually transformed into Chinese by adopting the phonetic approach which simulates their pronunciation with respect to Chinese syllabic structures (Lin, 2007, pp. 235-239), e.g. 扎克伯格 (*Zhākèbógé*, 'Zuckerberg'), 捷克斯洛伐克(Jiékèsīluòfákè, '*Czechoslovakia*)'.[136] When selecting Chinese characters for adaptation, the phonetic criterion might reduce or even override the influence of the law (cf. Chapter 1.4). UD annotates proper nouns with a special part-of-speech category (i.e. `PROPN`). However, their origin is not distinguished and the annotation also includes Chinese proper nouns (e.g. 中国, *Zhōngguó*, 'China' or 北京, *Běijīng*, 'Peking'). Despite this drawback, we test their exclusion on both – the tokens and the types. The results are shown in Table 41.

Table 41. The number of words $n(W)$, the parameters (a, b, c) and the coefficient of determination R^2 of both the model (M1, M2) obtained by the analysis of word tokens and word types while measuring the character in components and excluding proper nouns. $R^2 \geq 0.90$ for the increase in constituent lengths of the shortest constructs or the overall increasing trend in constituent lengths is highlighted in grey.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	word tokens excluding proper nouns									
n(W)	4247		16483		8250		8233		72359	
a	2.07	2.82	1.87	2.37	1.90	2.27	1.83	2.54	1.91	2.06
b	-0.02	0.61	0.14	0.61	0.14	0.48	0.13	0.79	0.07	0.15
c		0.32		0.24		0.17		0.33		0.05
R^2	0.0337	0.9696	0.3342	>0.99	0.5684	0.9894	0.0477	0.9856	0.0031	0.1463
	word types excluding proper nouns									
n(W)	763		4145		2589		2522		12015	
a	2.17	2.64	2.24	2.50	2.20	2.36	2.18	2.68	2.42	2.33
b	-0.04	0.31	-0.01	0.21	0.03	0.16	-0.02	0.39	-0.08	-0.19
c		0.19		0.11		0.07		0.21		-0.05
R^2	0.3620	>0.99	0.0531	0.9442	0.1458	0.9976	0.0803	0.9307	0.6955	0.7885

Neither the exclusion of the proper nouns corroborates the hypothesis. The goodness-of-fit is highly unsatisfactory, which applies to both – the tokens and the types. However, most samples yield negative values of the parameters b of M1 when the types without proper nouns

[136] Foreign words are also adapted in Chinese based on the semantic approach, which either opts for characters relating to the meaning of a foreign word or translates each morpheme (Lin, 2007, pp. 235-239).

are analysed in contrast to the tokens without proper nouns or even the previous results of the types (Table 40). The changes in the values of the parameters b indicate a positive impact of the exclusion on the results. For example, in the case of GSD, the types without proper nouns show $b = -0.08$ and $R^2 = 0.6955$ compared to $b = 0.07$ and $R^2 = 0.0031$ yielded by the tokens without proper nouns, and $b = -0.04$ and $R^2 = 0.3960$ yielded by all types.

Secondly, as mentioned in Chapter 3.2.5, the decomposition of Chinese characters into components depends on a given approach which might have an impact on MCL and the results. Hence, we also use an open-source document released by the Character Information Service Environment project (CHISE: CHaracter Information Service Environment, 2021). The document contains 20,951 Chinese characters, which are decomposed into components (IDS UCS Basic, 2022) while using Ideographic Description Characters, called Unicode blocks, e.g. ⿰ or ⿱ (Unicode, 2021). We apply this alternative approach (termed as CHISE) not only to a) all word tokens and word types but also to b) their selections which exclude proper nouns. However, we present within this thesis only the a) results (see Table 42) because we cannot conclude what precisely influences the results when the proper nouns are left out of the analysis. As addressed above, the exclusion also affects Chinese proper nouns. Nevertheless, both the results – a) and b) – are available on Github.[137]

Table 42. The number of words $n(W)$, the parameters (a, b, c) and the coefficient of determination R^2 of both the models (M1, M2) obtained by the analysis of word tokens and word types while using the CHISE decomposition of Chinese characters. $R^2 \geq 0.90$ for the increase in constituent lengths of the shortest constructs or the overall increasing trend in constituent lengths is highlighted in grey.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	word tokens – CHISE decomposition									
n(W)	4304		17845		8699		9146		81058	
a	1.74	2.17	1.72	1.78	1.74	1.81	1.71	1.77	1.72	1.74
b	-0.003	0.43	0.04	0.10	0.05	0.14	0.04	0.11	0.03	0.01
c		0.22		0.03		0.04		0.03		0.00
R²	NA	>0.99	0.3369	0.7638	0.5543	0.9806	0.2022	0.7670	0.4638	0.5213
	word types – CHISE decomposition									
n(W)	779		4944		2876		3082		15870	
a	1.84	2.05	1.84	1.85	1.83	1.87	1.82	1.84	1.92	1.88
b	-0.04	0.15	0.001	0.01	0.01	0.06	0.004	0.03	-0.03	-0.10
c		0.11		0.004		0.02		0.01		-0.02
R²	0.6563	>0.99	NA	0.3273	0.1045	0.5502	NA	0.5656	0.2267	0.9468

The alternative decomposition of Chinese characters does not bring about any considerable changes in the results. The fit between the model and the data remains poor and

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

the hypothesis is rejected for both – the tokens and the types. Only GSD types fitted by M2 show $R^2 \geq 0.90$. GSD is the only sample where $MCLs$ decrease along with the increase in WLs except for the last two $MCLs$. While M2 fits the increase, the fit of M1 is poor. As for HK-P types and M2, $R^2 \geq 0.99$ results from the three parameters being fitted to three word lengths after pooling the data.

Thirdly, we use the BLCU and CHISE sources for a maximal decomposition of Chinese characters into their components. We decompose each character until all its identified components cannot be decomposed further in a given source. To take the Chinese character 影 (*yǐng*, 'shadow; image') as an example, the original decomposition based on the CHISE source decomposes the character into two components, i.e. it ends in the first round in Table 43. The maximal decomposition continues to decompose each component until it ends in the fourth round in Table 43, and the character eventually has five components.

Table 43. The example of the maximal decomposition of a Chinese character based on the CHISE source.

Round	Character	Components	Maximal decomposition (Y/N)
1st	影	景, 彡	N
2nd	景	日, 京	N
	彡	彡	**Y**
3rd	日	日	**Y**
	京	亠, 口, 小	N
4th	亠	亠	**Y**
	口	口	**Y**
	小	小	**Y**

We apply this approach to all word tokens and word types and present the results in Table 44. The results of word tokens and word types without the proper nouns are available on Github.[138]

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

Table 44. The number of words $n(W)$, the parameters (a, b, c) and the coefficients of determination R^2 of both the models (M1, M2) obtained by the analysis of word tokens and word types based on the maximal BLCU and CHISE decomposition. $R^2 \geq 0.90$ for the increase in constituent lengths of the shortest constructs or the overall increasing trend in constituent lengths is highlighted in grey.

	HK-P		PUD		PUD-N		PUD-W		GSD	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	word tokens – BLCU maximal decomposition									
n(W)	4303		17844		8699		9145		80978	
a	2.20	2.92	1.98	2.21	2.02	2.29	1.94	2.16	2.05	2.18
b	0.07	0.66	0.17	0.41	0.19	0.45	0.18	0.40	0.14	0.22
c		0.29		0.10		0.12		0.09		0.04
R^2	NA	0.9426	0.5594	0.9715	0.6787	0.9789	0.6121	0.9631	0.6377	0.8352
	word tokens – CHISE maximal decomposition									
n(W)	4304		17845		8699		9146		81058	
a	2.07	2.89	2.04	2.14	2.07	2.23	2.01	2.13	2.04	2.08
b	-0.02	0.66	0.06	0.15	0.05	0.22	0.07	0.21	0.05	0.06
c		0.34		0.04		0.07		0.06		0.01
R^2	0.0560	0.9085	0.3159	0.7845	0.1609	0.9434	0.2009	0.9555	0.3938	0.4489
	word types – BLCU maximal decomposition									
n(W)	778		4943		2876		3081		15815	
a	2.41	2.60	2.48	2.53	2.43	2.51	2.39	2.44	2.81	2.69
b	0.01	0.14	0.03	0.08	0.05	0.12	0.05	0.12	-0.05	-0.13
c		0.07		0.02		0.03		0.03		-0.03
R^2	NA	>0.99	0.2312	0.4809	0.6129	0.8875	0.4993	0.6426	NA	0.6275
	word types – CHISE maximal decomposition									
n(W)	779		4944		2876		3082		15870	
a	2.19	2.42	2.24	2.28	2.21	2.34	2.16	2.25	2.51	2.41
b	-0.03	0.15	0.002	0.04	0.01	0.14	0.02	0.11	-0.07	-0.16
c		0.10		0.02		0.06		0.04		-0.03
R^2	0.5384	>0.99	NA	0.4891	NA	0.8360	NA	0.8380	0.4858	0.8590

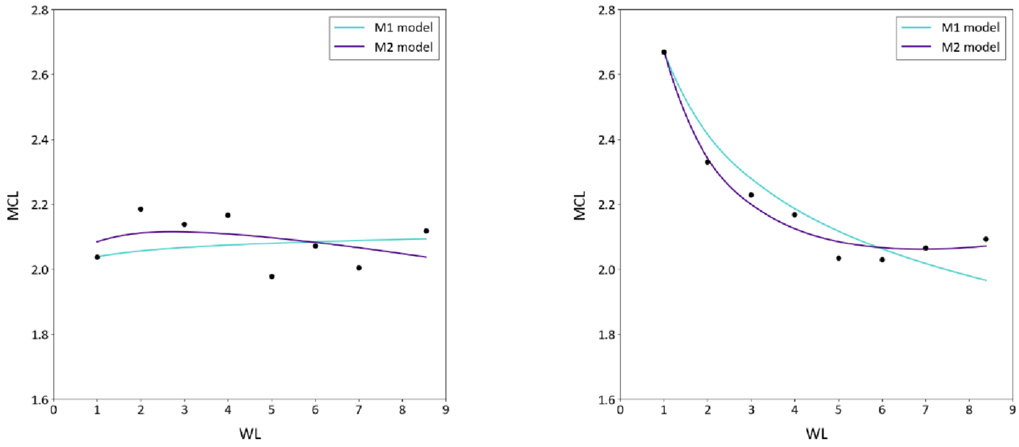
The maximal BLCU and CHISE decompositions show a similar trend in results when compared with their initial approach (for BLCU in Table 39 and Table 40, and for CHISE in Table 42). The parameter a of M1 obviously reaches higher values due to the increase in the number of components, however, the parameter b of M1 increases too. Its positive values contradicting the law prevail and the hypothesis is not corroborated for both – the tokens and the types.

The alternative approaches to the character length do not considerably change the results, which shifts focus towards the word length. The results obtained on the phrasal level showed that the prevalence of one- and two-character words in our samples and generally in Chinese does not prevent the law from coming into play when the Chinese word is the constituent. The questions arise whether this also applies to the Chinese word in the construct position and whether we deal with another factor influencing the results, e.g. the word segmentation. Therefore, we apply the law to an additional sample, i.e. the Lancaster Corpus of Mandarin Chinese, LCMC (McEnery, Xiao and Mo, 2003), tested also by studies on Chinese (e.g. on the sentence level by Hou et al., 2017; on the clause level by Hou et al., 2019a; on the word level by Chen and Liu, 2022). The corpus contains more than 800k word tokens and the word segmentation was performed using the Chinese Lexical Analysis System ICTCLAS (Institute of Computing Technology of Chinese Academy of Science, n.d.).[139] We test the LCMC sample on word tokens and word types consisting only of Chinese characters, which we decompose while using the BLCU source. The obtained results are shown in Table 45 and Figure 35.

Table 45. MAL applied to the triplet of the word, character and component – tokens and types from LCMC.

LCMC: tokens – BLCU			LCMC: types – BLCU		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	399070	2.04	1	3515	2.67
2	390050	2.19	2	28263	2.33
3	24882	2.14	3	5976	2.23
4	12727	2.17	4	4385	2.17
5	614	1.98	5	224	2.03
6	133	2.07	6	88	2.03
7	125	2.00	7	37	2.07
8.54	24	2.12	8.39	18	2.09
	M1	M2		M1	M2
a	2.04	2.11	*a*	2.67	2.58
b	0.01	0.04	*b*	-0.14	-0.24
c		0.01	*c*		-0.03
R^2	NA	0.1158	R^2	0.8866	0.9771

[139] ICTCLAS software is also commonly used for the word segmentation of samples compiled by authors of studies on Chinese, e.g. by Chen and Liu (2016), Hou et al. (2017), Jin and Liu (2017) and Hou et al. (2019a).



LCMC – tokens

LCMC – types

Figure 35. MAL applied to the triplet of word, character and component – tokens and types from LCMC.

In the case of the word tokens, mean character lengths show the zig-zag tendency with MCL_2 being the highest. The fit between the models and the data is highly unsatisfactory and the hypothesis is not corroborated. When analysing the types, we yield completely different results. The overall trend in $MCLs$ is decreasing. The hypothesis is not rejected for M2 which R^2 reaches the standard of $R^2 \geq 0.90$. R^2 of M1 is lower, i.e. $R^2 = 0.8866$ while $b = -0.14$. The value might be influenced by the second regime which occurs with the last two $MCLs$ belonging to words length equal to or greater than seven characters (highlighted in yellow in Table 45). Firstly, these words have an extremely low relative frequency ($0.13\ \%$). If we test their omission, the fit of M1 increases from 0.8866 to 0.9792. Secondly, these words represent common nouns (e.g. 中国人民解放军, *Zhōngguó Rénmín Jiěfàngjūn*, 'Chinese People's Liberation Army'), numerals (e.g. 百分之九十九点九, *bǎi fēnzhī jiǔshíjiǔ diǎn jiǔ*, '99.9 %'), fixed expressions or idioms (e.g. 车到山前必有路, *chē dào shānqián bì yǒu lù*, 'Things will eventually sort themselves out.') and proper nouns (e.g. 布拉格斯巴达队, *Bùlāgé Sībādá duì*, 'Sparta Prague team'). Hence, the question arises whether the deviation of their $MCLs$ from the decreasing trend is caused by the low frequency or compound form (cf. Mačutek and Rovenchak, 2011, p. 139). UD, for example, segments the noun 中国人民解放军 (*Zhōngguó rénmín jiěfàngjūn*, 'Chinese People's Liberation Army') into four separate words, i.e. 1) 中国 (*Zhōngguó*, 'China'), 2) 人民 (*rénmín*, 'the people'), 3) 解放 (*jiěfàng*, 'to liberate, liberation'), 4) 军 (*jūn*, 'army'). Based on all these results from LCMC, we can draw two preliminary conclusions – firstly, the law appears to be highly sensitive to word segmentation. Secondly, the law comes into force even on the word level if the frequency of usage is not taken into account. Regarding short-term memory, none of the word or character lengths exceeds its upper threshold (i.e. 7 ± 2, Miller, 1956).

As in the case of the UD samples, we apply all the alternative approaches to LCMC, i.e. the exclusion of proper nouns distinguished in LCMC by special part-of-speech categories[140], the CHISE decomposition and the maximal decomposition based on the BLCU and the CHISE sources. Regarding the alternative decompositions, we again test both – a) all word tokens and word types and b) their selections which exclude proper nouns. Since we cannot again conclude what precisely influences the results because the exclusion includes foreign and Chinese proper nouns, we present only the a) results (Table 46) while making both available on Github.[141]

Table 46. The number of words $n(W)$, the parameters (a, b, c) and the coefficients of determination R^2 of both the models (M1, M2) obtained by the analysis of word tokens and types while excluding proper nouns and using different decompositions of Chinese characters – LCMC.

	LCMC tokens		LCMC types	
	M1	M2	M1	M2
	excluding proper nouns			
n(W)	791644		37795	
a	2.03	2.12	2.66	2.63
b	0.0001	0.04	-0.17	-0.24
c		0.02		-0.02
R^2	NA	0.1509	0.8743	0.8895
	CHISE decomposition			
n(W)	827726		42543	
a	1.78	1.83	1.96	1.96
b	-0.03	-0.003	-0.09	-0.11
c		0.01		-0.01
R^2	0.5792	0.7212	0.9316	0.9371
	BLCU maximal decomposition			
a	2.19	2.31	3.05	2.94
b	0.02	0.07	-0.17	-0.28
c		0.03		-0.04
R^2	NA	0.2169	0.9118	0.9812
	CHISE maximal decomposition			
a	2.15	2.21	2.71	2.64
b	-0.05	-0.02	-0.18	-0.25
c		0.01		-0.02
R^2	0.6700	0.7588	0.9472	0.9714

[140] I.e. `nr` for personal names, `ns` for place names, `nt` for organization names and `nz` for other proper nouns.

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

Regardless of the alternative approach, the tokens yield considerably worse fitting results than the types – the fit of M1 is either not available (the parameter b of M1 has positive values) or unsatisfactory, as in the case of M2. Hence, the hypothesis is not corroborated. As for the types, the hypothesis is rejected with respect to the standard of $R^2 \geq 0.90$ only when we exclude the proper nous. Nevertheless, the parameter b of M1 has a negative value and values of R^2 of both the models are high compared to the tokens. The law again manifests itself differently depending on the analysis of the tokens or the types. Additionally, the decomposing approach to characters can be regarded as another decisive factor when evaluating the results of the types by optics of the standard of $R^2 \geq 0.90$. The initial decomposition based on the BLCU source yielded a fit of M1 which rejected the hypothesis ($R^2 = 0.8866$, Table 45). Despite the word lengths equal to or greater than seven characters which keep deviating from the decreasing trend, the hypothesis becomes corroborated when the maximal BLCU decomposition is applied ($R^2 = 0.9118$ for M1). R^2 of the CHISE decomposition reaches even a higher value ($R^2 = 0.9316$) and R^2 of its maximal decomposition is the highest ($R^2 = 0.9472$).

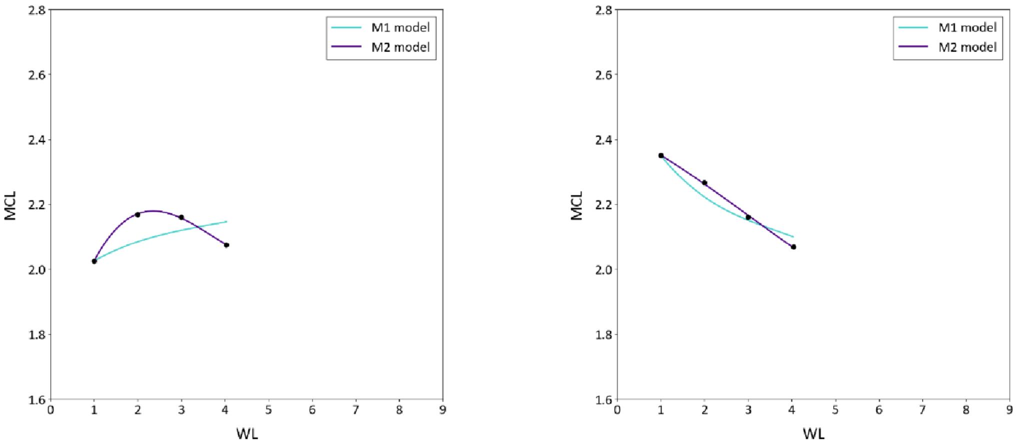
However, an objection can be raised that the size of LCMC, which is considerably larger than the size of HK-P, PUD and GSD, contributes to or lead to these results. For this reason, we test tokens and types from an LCMC collection of sci-fi texts (i.e. LCMC:M) which contains 10,054 tokens and 2,803 types. [142,143] In both cases, we use words consisting only of Chinese characters, which we decompose based on the BLCU source. The results are included in Table 47 and Figure 36.

Table 47. MAL applied to the triplet of the word, character and component – tokens and types from LCMC:M.

LCMC:M tokens			LCMC:M types		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	5064	2.03	1	545	2.35
2	4660	2.17	2	2038	2.27
3	241	2.16	3	139	2.16
4.03	89	2.07	4.04	81	2.07
	M1	M2		M1	M2
a	2.03	2.27	*a*	2.35	2.47
b	0.04	0.26	*b*	-0.08	0.02
c		0.11	*c*		0.05
R²	0.0368	0.9990	*R²*	0.9343	0.9987

[142] We do not aim in the scope of the thesis to test differences in results of various LCMC text types. The goal is to test only the potential impact of the sample size.

[143] For comparison, PUD-W consists of 9,145 word tokens and 3,081 word types.



LCMC:M – tokens

LCMC:M – types

Figure 36. MAL applied to the triplet of the word, character and component – tokens and types from LCMC:M.

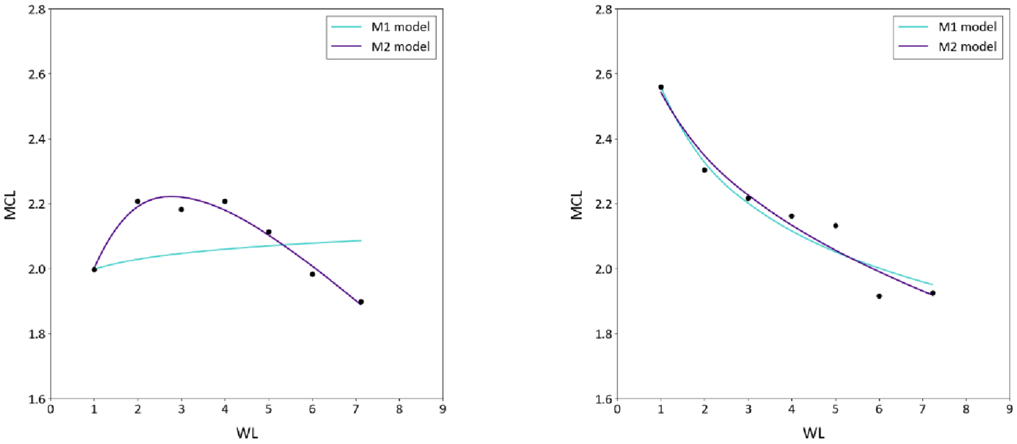
When comparing the results yielded by M1 (Table 45 and Table 47), both the samples – LCMC and LCMC:M – show a similar trend. The tokens yield poor goodness-of-fit and reject the hypothesis (LCMC, LCMC:M), while the types do not (LCMC:M) or their $MCLs$ show apparent menzerathian decreasing tendency and R^2 approximating the standard of $R^2 \geq 0.90$ (LCMC). Despite the different size of the samples, the word segmentation and the analysis of unit types are still crucial factors for the law to come into play or the menzerathian decreasing tendency to appear. However, the LCMC:M types show a fit of M1 which corroborates the hypothesis ($R^2 = 0.9343$), whereas the LCMC types do not ($R^2 = 0.8866$, Table 45). Firstly, LCMC:M does not include the word lengths equal to or greater than seven characters which deviate from the decreasing tendency and contribute to the worse fitting result in LCMC. Secondly, the sample contains only sci-fi texts, while LCMC includes texts of 15 different text types. Hence, the question arises whether we can consider the sample homogeneity rather than the sample size as another factor to which the law positively responds.

In addition, Chen and Liu (2022) applied the law to a sample which merges two LCMC text collections, i.e. press reportages (LCMC:A) and academic prose (LCMC:J). The authors fitted obtained WLs and $MCLs$ with the complete model and yielded an extremely low fit, i.e. $R^2 = 0.1625$. However, the word tokens were analysed. Despite differences in the approach[144], we also run the analysis for this merged sample (labelled as LCMC:A+J) which contains 205,649 tokens and 19,294 types. We use only Chinese words whose characters are decomposed based on the BLCU source. The results are shown in Table 48 and Figure 37.

[144] E.g. we do not use the same source for decomposing Chinese characters into components as Chen and Liu (2022).

Table 48. MAL applied to the triplet of the word, character and component – tokens and types from LCMC:A+J.

LCMC:A+J tokens			LCMC:A+J types		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	86434	2.00	1	2280	2.56
2	108327	2.21	2	13051	2.30
3	7407	2.18	3	2436	2.22
4	3265	2.21	4	1420	2.16
5	161	2.11	5	74	2.13
6	30	1.98	6	20	1.92
7.12	25	1.90	7.23	13	1.93
	M1	M2		M1	M2
a	2.00	2.20	*a*	2.56	2.58
b	0.02	0.27	*b*	-0.14	-0.09
c		0.10	*c*		0.02
R^2	NA	0.9647	R^2	0.9422	0.9519



LCMC:A+J – tokens

LCMC:A+J – types

Figure 37. MAL applied to the triplet of the word, character and component – tokens and types from LCMC:A+J.

The tokens show unsatisfactory goodness-of-fit and reject the hypothesis, which corresponds to the results yielded by Chen and Liu (2022). R^2 of M1 is not available (the parameter b of M1 is positive) and M2 fits the second regime of $MCLs$ (highlighted in yellow in Table 48). As regards the types, the hypothesis is not rejected. R^2 of both the models meets the standard of $R^2 \geq 0.90$ while $MCLs$ decrease. If we compare the LCMC samples with regard to the fit of M1, the results obtained from LCMC:A+J approximate the results from LCMC:M. Firstly, LCMC:A+J includes words having lengths equal to or greater than seven characters to a lesser extent than LCMC. Secondly, even though LCMC:A+J merges two text collections (press reportages and academic prose), its homogeneity might be higher than in LCMC. Apart from the word segmentation and the frequency of usage, we might deal with the influence of sample homogeneity (or heterogeneity). Finally, it is noteworthy that Chen and Liu (2022) left the Chinese character out of the unit hierarchy and used the component instead. Based on the results obtained by the analysis of word tokens, the authors concluded that the triplet of the word, component and stroke is the only accepted unit combination in written Chinese. However, our results show that the Chinese character is the direct constituent of the word when the types are analysed.

Finally, we present results of LCMC:M and LCMC:A+J when alternative approaches are applied, i.e. the exclusion of the proper names, the CHISE decomposition and the maximal decomposition based on both the sources (BLCU and CHISE). The result of all word tokens and word types are presented in Table 49. The results obtained by the alternative decomposing approaches applied to word tokens and word types without proper nouns are available on Github.[145]

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

Table 49. The number of words $n(W)$, the parameters (a, b, c) and the coefficients of determination R^2 of both the models (M1, M2) obtained by the analysis of word tokens and types while excluding proper nouns and using different decompositions of Chinese characters – LCMC:M and LCMC:A+J. $R^2 \geq 0.90$ for the increase in constituent lengths of the shortest constructs or the overall increasing trend in constituent lengths is highlighted in grey.

	LCMC:M		LCMC:M		LCMC:A+J		LCMC:A+J	
	tokens		types		tokens		types	
	M1	M2	M1	M2	M1	M2	M1	M2
	excluding proper nouns – BLCU							
n(W)	9612		2732		199164		17623	
a	2.01	2.27	2.35	2.50	1.99	2.23	2.55	2.63
b	0.05	0.29	-0.08	0.04	0.004	0.26	-0.16	-0.10
c		0.12		0.06		0.10		0.03
R²	0.0901	0.9982	0.9071	0.9997	NA	0.7693	0.8609	0.8799
	CHISE decomposition							
n(W)	10055		2804		205657		19300	
a	1.78	1.95	1.88	1.97	1.80	1.90	1.94	1.99
b	-0.003	0.17	-0.05	0.04	-0.03	0.12	-0.09	-0.02
c		0.09		0.05		0.06		0.03
R²	0.0144	0.9933	0.8876	0.9989	0.4207	0.9427	0.8844	0.9475
	BLCU maximal decomposition							
a	2.15	2.50	2.63	2.76	2.14	2.41	2.91	2.94
b	0.08	0.37	-0.09	0.001	0.03	0.33	-0.16	-0.11
c		0.15		0.05		0.12		0.02
R²	0.3089	0.9974	0.9520	0.9993	NA	0.9470	0.9484	0.9575
	CHISE maximal decomposition							
a	2.12	2.32	2.37	2.49	2.17	2.32	2.59	2.62
b	-0.003	0.18	-0.10	0.003	-0.06	0.13	-0.17	-0.12
c		0.09		0.05		0.07		0.02
R²	0.0127	0.9014	0.9511	0.9990	0.5554	0.9761	0.9515	0.9595

Both the samples yield similar results as LCMC regarding the tokens – the hypothesis is not corroborated. R^2 of M1 is unsatisfactory, i.e. having low values or positive parameters b. Although R^2 of M2 reaches values higher than 0.90, the model fits the second regime. In the case of the types, slight differences in the goodness-of-fit appear. Firstly, LCMC:M corroborates the hypothesis for the excluded proper nouns, LCMC:A+J does not as LCMC. Secondly, only LCMC does not reject the hypothesis when M1 and the CHISE decomposition are applied, whereas LCMC:M and LCMC:A+J do ($R^2 = 0.8876$ and $R^2 = 0.8844$ accordingly). Thirdly, both the maximal decompositions (BLCU, CHISE) always yield $R^2 \geq 0.90$ without regard to the model (M1, M2) or the sample (LCMC, LCMC:M, LCMC:A+J).

To sum it up, the triplet of the word, character and component shows that several factors influence the results. First and foremost, the law is highly sensitive to word segmentation and unit frequency, as demonstrated by the difference between the UD and LCMC samples. In addition, the trends in the results, which are not the same across the LCMC samples, also indicate that the law might respond to the degree of sample homogeneity (or heterogeneity) and a decomposing approach to Chinese characters.

4.4.2 The character and stroke as constituents

Hypothesis: the longer the word length measured in the number of Chinese characters, the shorter the mean length of the characters measured in strokes.

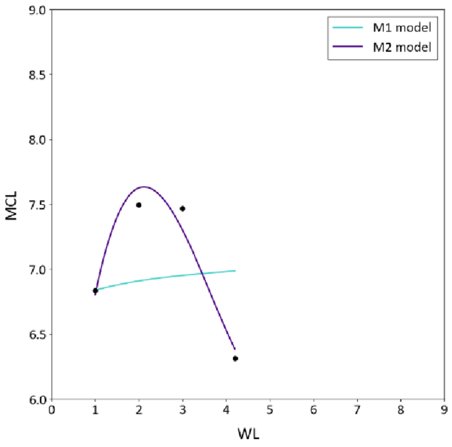
We present the results of word tokens and word types in Table 50 and Table 51 respectively, and Figure 38. WL denotes the word length measured the in the number of Chinese characters, $f(WL)$ its frequency and MCL the mean character length measured in the number of strokes. Both the models are applied – M1 labelling the truncated model $y(x) = ax^b$ and M2 labelling the complete model $y(x) = ax^b e^{cx}$. Their parameters (a, b, c) and the coefficient of determination R^2 are shown in the tables. MCL_1 is used as the parameter a when M1 is fitted to the data. Finally, if a value of MCL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 50. MAL applied to the triplet of the word, character and stroke – tokens.

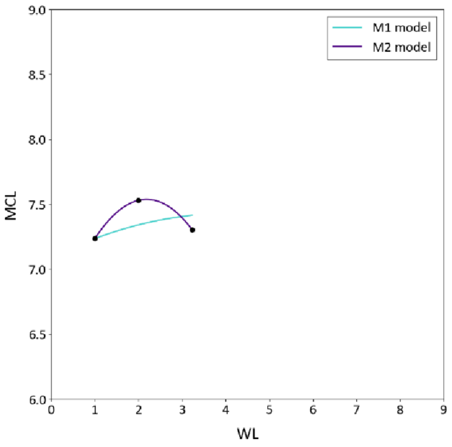
HK-P			PUD			PUD-N			PUD-W			GSD		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	1989	6.84	1	7784	6.31	1	3946	6.42	1	3838	6.20	1	36504	6.54
2	2155	7.50	2	8683	7.52	2	4195	7.59	2	4488	7.45	2	42058	7.59
3	149	7.47	3	1020	7.84	3	441	7.89	3	579	7.79	3	1538	7.95
4.20	10	6.32	4	274	7.97	4	94	8.14	4	180	7.87	4	707	7.46
			5	67	7.97	5.26	23	7.95	5	49	7.98	5	126	8.05
			6.25	16	7.88				6.27	11	7.82	6	30	8.29
												8	15	8.18
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	6.84	8.87	*a*	6.31	6.86	*a*	6.42	7.03	*a*	6.20	6.76	*a*	6.54	6.82
b	0.02	0.55	*b*	0.15	0.33	*b*	0.16	0.36	*b*	0.16	0.35	*b*	0.12	0.18
c		0.27	*c*		0.08	*c*		0.09	*c*		0.08	*c*		0.02
R^2	NA	0.9474	R^2	0.6882	0.9909	R^2	0.7776	0.9935	R^2	0.6859	0.9881	R^2	0.7166	0.8279

Table 51. MAL applied to the triplet of the word, character and stroke – types.

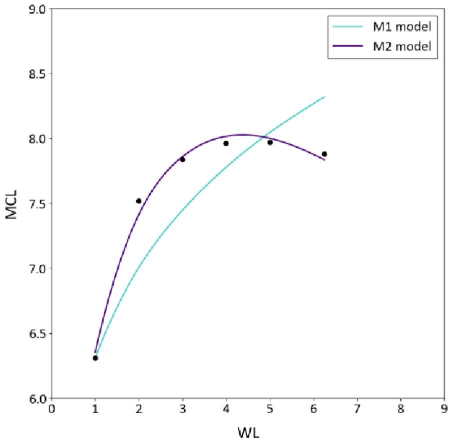
HK-P			PUD			PUD-N			PUD-W			GSD		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	198	7.24	1	469	7.91	1	339	7.70	1	332	7.41	1	1485	8.88
2	532	7.53	2	3456	7.93	2	2090	7.91	2	2138	7.80	2	12715	8.05
3.23	48	7.31	3	742	7.88	3	347	7.89	3	429	7.84	3	953	8.01
			4	207	7.97	4	77	8.13	4	135	7.86	4	509	7.83
			5	53	8.29	5.26	23	7.95	5	36	8.42	5	111	8.05
			6.25	16	7.88				6.27	11	7.82	6	27	8.44
												8	15	8.18
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	7.24	8.06	*a*	7.91	7.90	*a*	7.70	7.80	*a*	7.41	7.56	*a*	8.88	8.46
b	0.02	0.21	*b*	0.01	0.01	*b*	0.03	0.06	*b*	0.05	0.10	*b*	-0.06	-0.16
c		0.11	*c*		0.001	*c*		0.01	*c*		0.02	*c*		-0.04
R^2	NA	>0.99	R^2	0.1071	0.1137	R^2	0.6278	0.7197	R^2	0.4882	0.5786	R^2	NA	0.7100



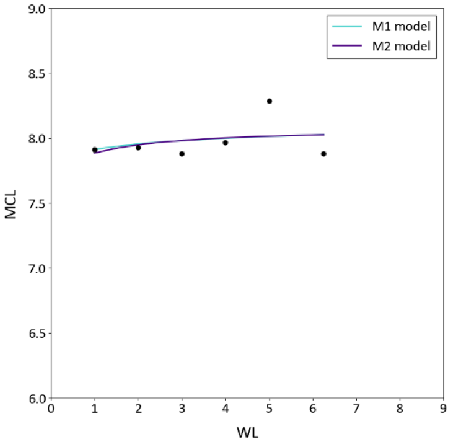
HK-P – tokens



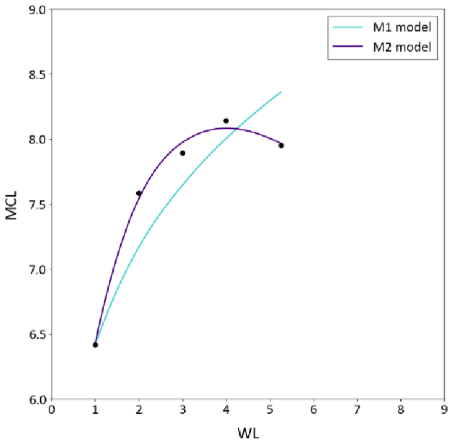
HK-P – types



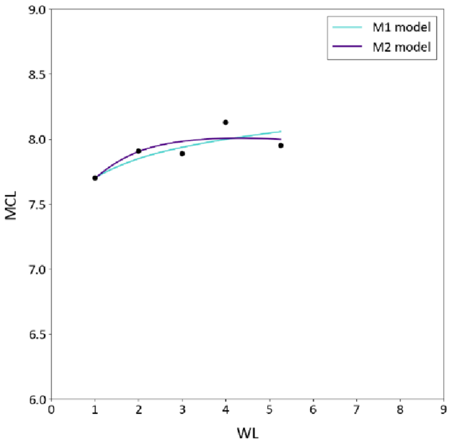
PUD – tokens



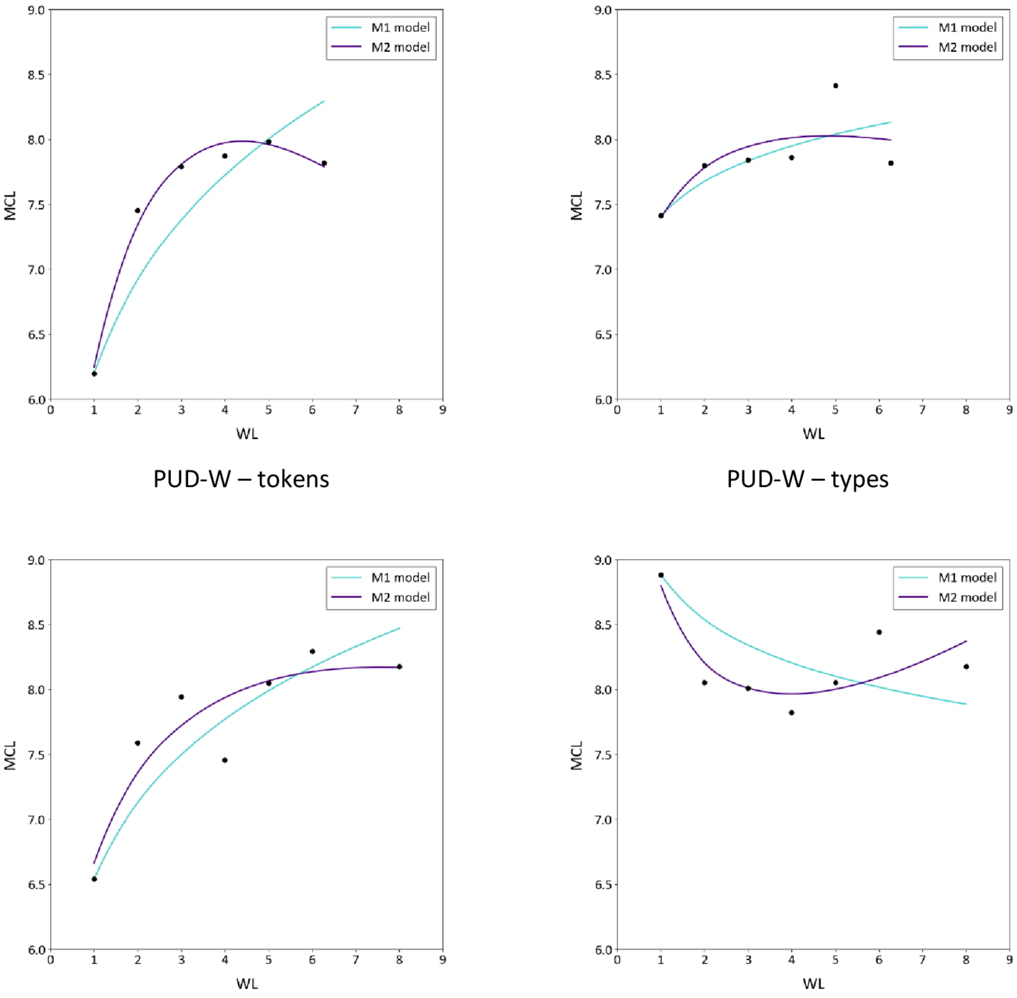
PUD – types



PUD-N – tokens



PUD-N – types



PUD-W – tokens

PUD-W – types

GSD – tokens

GSD – types

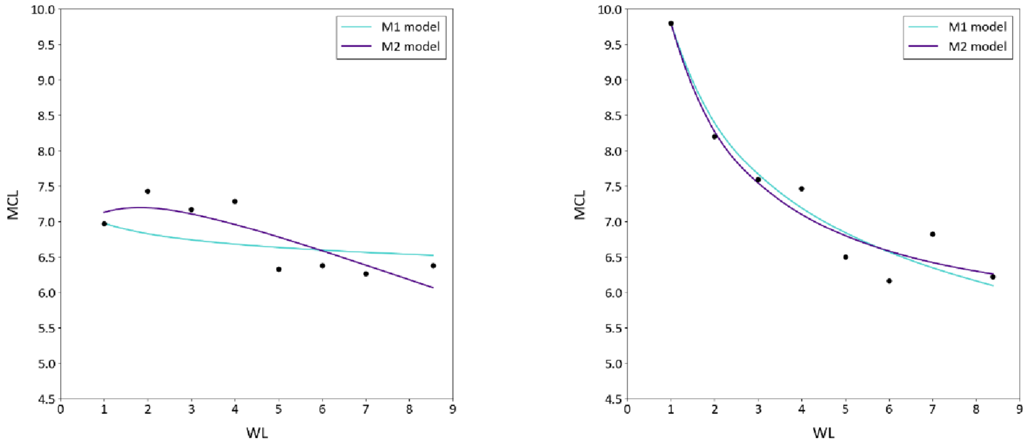
Figure 38. MAL applied to the triplet of the word, character and stroke – tokens and types.

Opting for the stroke as the measurement unit for the Chinese character yields similar results as in the case of the component. Neither the tokens nor the types corroborate the hypothesis. When analysing the tokens, mean character lengths contradict the law. MCL_1s are lower than MCL_2 or the lowest and the rest of the mean character lengths continue increasing along with WL until they peak with second to the last $MCLs$ (except for HK-P). As for M1, all parameters b have positive values and R^2 is not available or is unsatisfactory. Although R^2 of M2 reaches values higher than 0.90, the model fits the increase in $MCLs$. When it comes to the types, values of MCL_1 in the PUD samples are still lower or the lowest while values of second to the last MCL are the highest. The parameter b of M1 remains positive in these samples. The only exception is GSD. The parameter b has a negative value and the fitting curve decreases from its head (MCL_1 is the highest). Nevertheless, its tail keeps rising and the result is unsatisfactory. In the case of HK-P, we do not evaluate the results of its types because they have only three word lengths after pooling the data (in the case of M2, $R^2 > 0.99$ because of the fit by three parameters). When taking the short-term memory span into account, all the samples show WL and MCL following its upper threshold (i.e. 7 ± 2, Miller, 1956).

We also apply the law to LCMC, LCMC:M and LCMC:A+J to test whether the stroke is an inappropriate measurement unit for the character or whether the UD word segmentation influences the results, as in the case of the component. Table 52 and Figure 39 present the results of LCMC, and Table 53 and Figure 40 show the results of LCMC:M and LCMC:A+J.

Table 52. MAL applied to the triplet of the word, character and stroke – tokens and types from LCMC.

LCMC: tokens			LCMC: types		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	399070	6.97	1	3515	9.80
2	390050	7.43	2	28263	8.20
3	24882	7.17	3	5976	7.59
4	12727	7.29	4	4385	7.47
5	614	6.33	5	224	6.50
6	133	6.38	6	88	6.16
7	125	6.26	7	37	6.82
8.54	24	6.38	8.39	18	6.22
	M1	M2		M1	M2
a	6.97	7.44	*a*	9.80	9.64
b	-0.03	0.07	*b*	-0.22	-0.27
c		0.04	*c*		-0.02
R^2	0.2960	0.6671	R^2	0.9387	0.9457



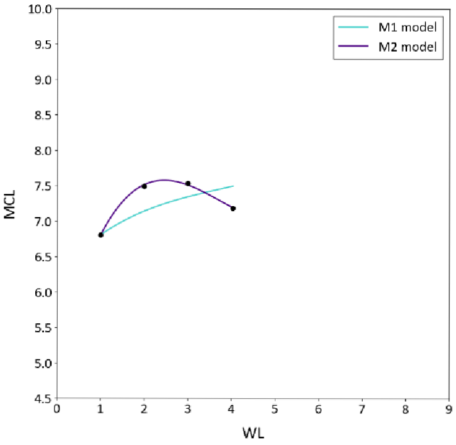
LCMC – tokens

LCMC – types

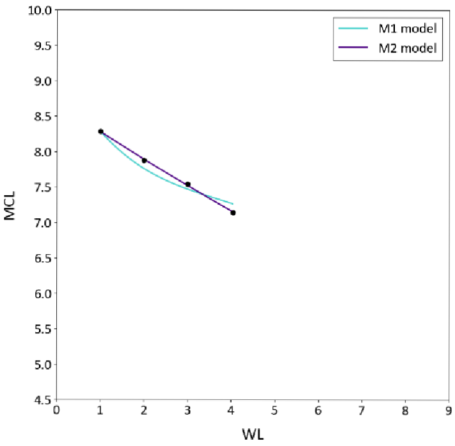
Figure 39. MAL applied to the triplet of the word, character and stroke – tokens and types from LCMC.

Table 53. MAL applied to the triplet of the word, character and stroke – tokens and types from LCMC:M and LCMC:A+J.

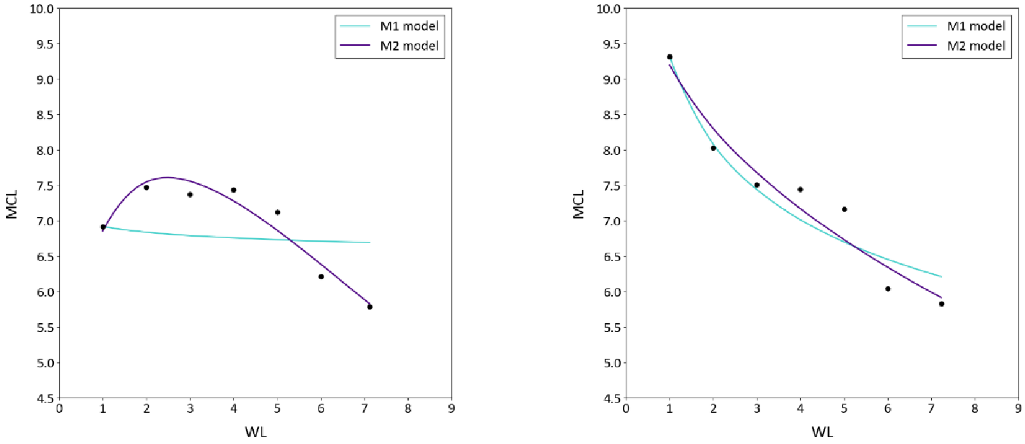
LCMC:M tokens			LCMC:M types			LCMC:A+J tokens			LCMC:A+J types		
WL	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*	*WL*	*f(WL)*	*MCL*
1	5064	6.81	1	545	8.28	1	86434	6.92	1	2280	9.32
2	4660	7.50	2	2038	7.88	2	108327	7.47	2	13051	8.03
3	241	7.54	3	139	7.54	3	7407	7.37	3	2436	7.51
4.03	89	7.18	4.04	81	7.15	4	3265	7.43	4	1420	7.45
						5	161	7.12	5	74	7.17
						6	30	6.21	6	20	6.04
						7.12	25	5.79	7.23	13	5.83
	M1	M2		M1	M2		M1	M2		M1	M2
a	6.81	7.84	*a*	8.28	8.69	*a*	6.92	7.82	*a*	9.32	9.61
b	0.07	0.35	*b*	-0.09	0.001	*b*	-0.02	0.33	*b*	-0.20	-0.08
c		0.14	*c*		0.05	*c*		0.13	*c*		0.04
R^2	0.2378	0.9969	R^2	0.9525	0.9990	R^2	0.0684	0.9373	R^2	0.9132	0.9428



LCMC:M – tokens



LCMC:M – types



LCMC:A+J – tokens

LCMC:A+J – types

Figure 40. MAL applied to the triplet of the word, character and stroke – tokens and types from LCMC:M and LCMC:A+J.

The goodness-of-fit between the models and the word tokens is unsatisfactory and the hypothesis is rejected. LCMC shows MCL_2 being the highest and the rest of $MCLs$ fluctuating in a zig-zag trend. In the case of LCMC:A+J, the peak of mean character lengths is also shifted towards MCL_2 and, in the case of LCMC:M, mean character lengths peak even with MCL_3. The analysis of the types brings changes in the results – $MCLs$ mostly decrease while WLs increase. R^2 of both the models meets the standard of $R^2 \geq 0.90$ in all three samples and the hypothesis is not rejected. Our results correspond to results yielded by Chen and Liu (2022), who tested LCMC:A+J on the same triplet while taking the frequency of usage into account – lengths of word tokens and mean character lengths measured in strokes did not corroborate the hypothesis ($R^2 = 0.5009$). All these results support the previous findings that the law is highly sensitive to word segmentation and unit frequency. Finally, considering short-term memory, $MCLs$ measured in strokes slightly exceed the upper threshold (i.e. 7 ± 2, Miller, 1956) when the types from LCMC and LCMC:A+J are analysed.

4.4.3 The syllable and sound as constituents

Hypothesis: the longer the word length measured in the number of syllables, the shorter the mean length of the syllables measured in sounds.

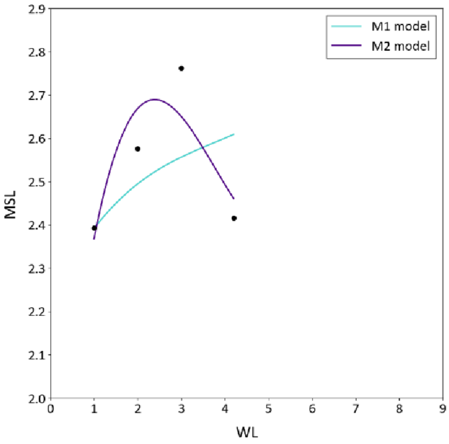
The results yielded when applying the law to word tokens are presented in Table 54. Table 55 presents the results of word types. Figure 41 visualises both the results. WL denotes the word length measured the in the number of syllables, $f(WL)$ its frequency and MSL the mean syllable length measured in the number of sounds. The data of the word tokens and types are fitted by the truncated model $y(x) = ax^b$ labelled as M1 and by the complete model $y(x) = ax^b e^{cx}$ labelled as M2. Their parameters (a, b, c) and the coefficient of determination R^2 are shown in the tables. As for M1, we fit the data with MSL_1, i.e. the mean syllable length of mono-syllabic words. Finally, if a value of MSL is higher compared to its predecessor (=second regime), we highlight the respective cells in yellow.

Table 54. MAL applied to the triplet of the word, syllable and sound – tokens.

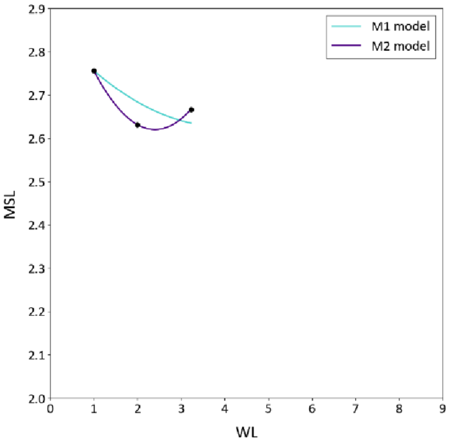
	HK-P			PUD			PUD-N			PUD-W			GSD	
WL	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*
1	1989	2.39	1	7785	2.34	1	3946	2.35	1	3839	2.33	1	36537	2.44
2	2155	2.58	2	8684	2.61	2	4195	2.61	2	4489	2.61	2	42109	2.64
3	150	2.76	3	1020	2.47	3	441	2.50	3	579	2.44	3	1543	2.32
4.20	10	2.42	4	274	2.28	4	94	2.33	4	180	2.25	4	709	2.39
			5	67	2.10	5.26	23	2.19	5	49	2.08	5	127	2.25
			6.25	16	2.22				6.27	11	2.21	6	30	2.17
												8	15	2.16
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.39	2.85	*a*	2.34	2.61	*a*	2.35	2.75	*a*	2.33	2.60	*a*	2.44	2.59
b	0.06	0.44	*b*	-0.01	0.18	*b*	0.005	0.34	*b*	-0.02	0.17	*b*	-0.04	0.04
c		0.18	*c*		0.09	*c*		0.15	*c*		0.09	*c*		0.04
R^2	0.0193	0.7319	R^2	0.0715	0.6113	R^2	NA	0.9383	R^2	0.0987	0.5955	R^2	0.4320	0.6935

Table 55. MAL applied to the triplet of the word, syllable and sound – types.

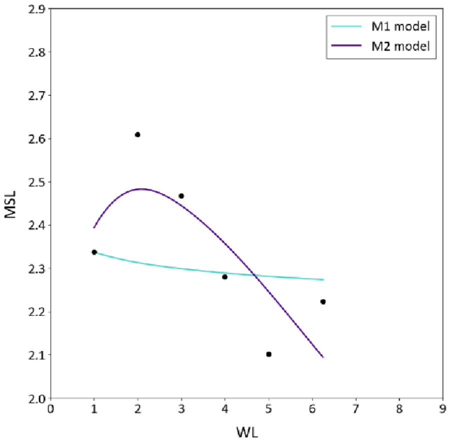
HK-P			PUD			PUD-N			PUD-W			GSD		
WL	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*
1	123	2.76	1	205	2.81	1	178	2.79	1	161	2.75	1	295	2.81
2	510	2.63	2	3078	2.67	2	1927	2.65	2	1970	2.66	2	9577	2.72
3.22	49	2.67	3	740	2.50	3	347	2.51	3	428	2.47	3	941	2.33
			4	206	2.30	4	76	2.36	4	135	2.27	4	502	2.34
			5	53	2.14	5.23	22	2.19	5	36	2.12	5	111	2.24
			6.27	15	2.23				6.27	11	2.21	6	27	2.17
												8	15	2.16
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	2.76	2.53	*a*	2.81	2.90	*a*	2.79	2.99	*a*	2.75	2.87	*a*	2.81	2.85
b	-0.04	-0.19	*b*	-0.14	-0.09	*b*	-0.12	0.03	*b*	-0.13	-0.06	*b*	-0.14	-0.15
c		-0.08	*c*		0.02	*c*		0.07	*c*		0.03	*c*		-0.003
R^2	0.5469	>0.99	R^2	0.8990	0.9218	R^2	0.9038	0.9997	R^2	0.8645	0.9024	R^2	0.9102	0.9154



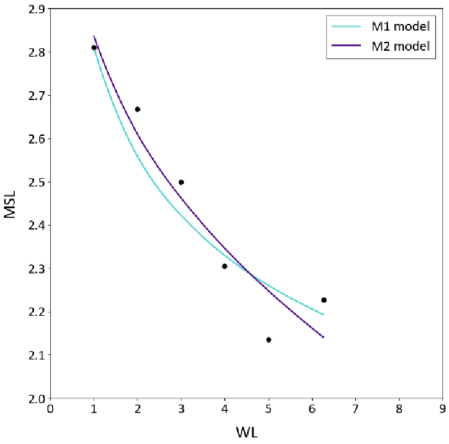
HK-P – tokens



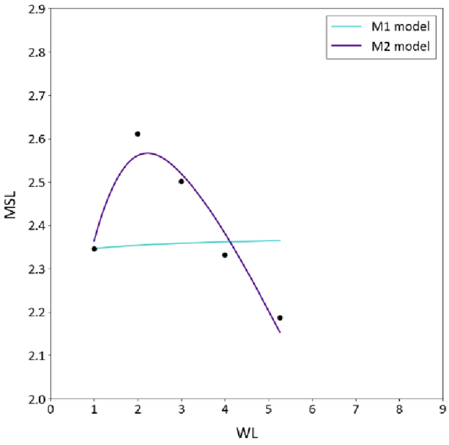
HK-P – types



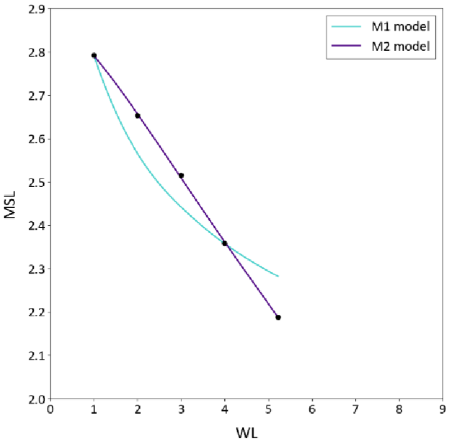
PUD – tokens



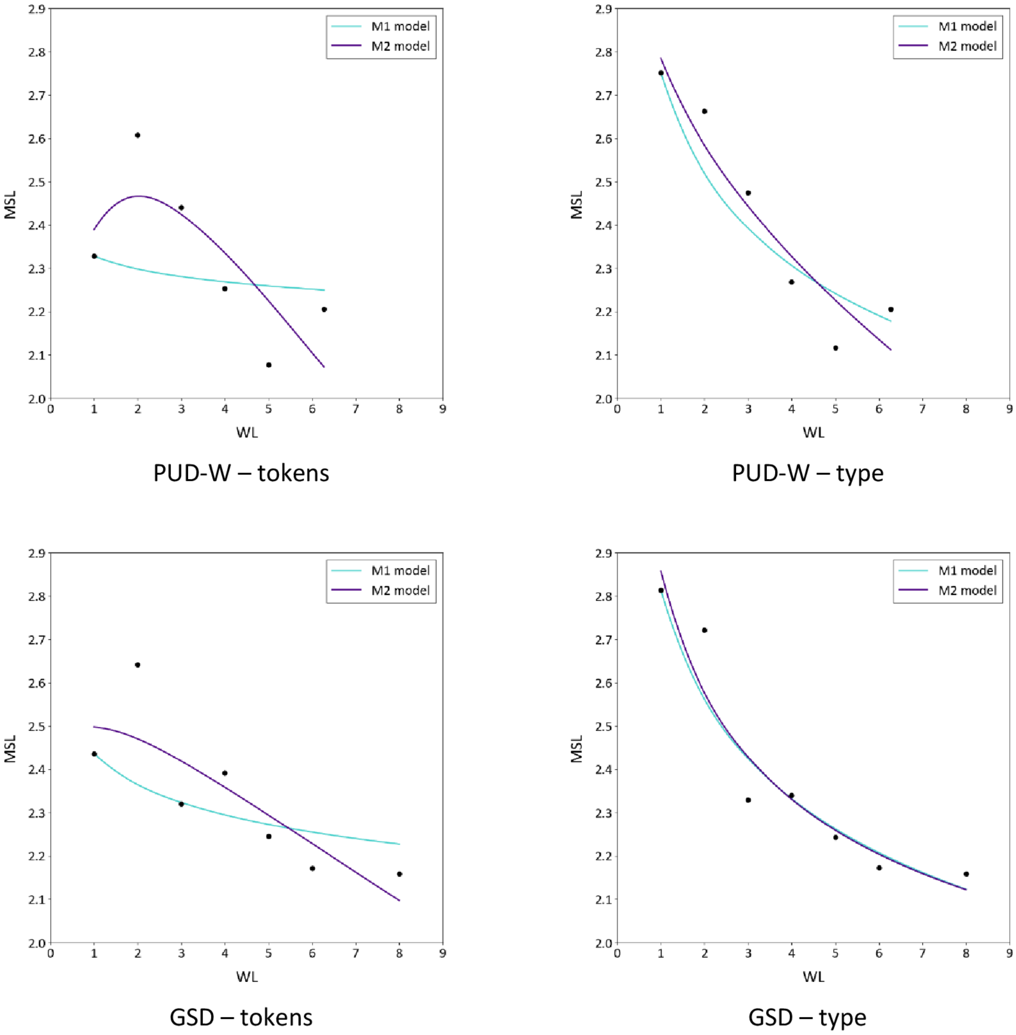
PUD – type



PUD-N – tokens



PUD-N – type



PUD-W – tokens

PUD-W – type

GSD – tokens

GSD – type

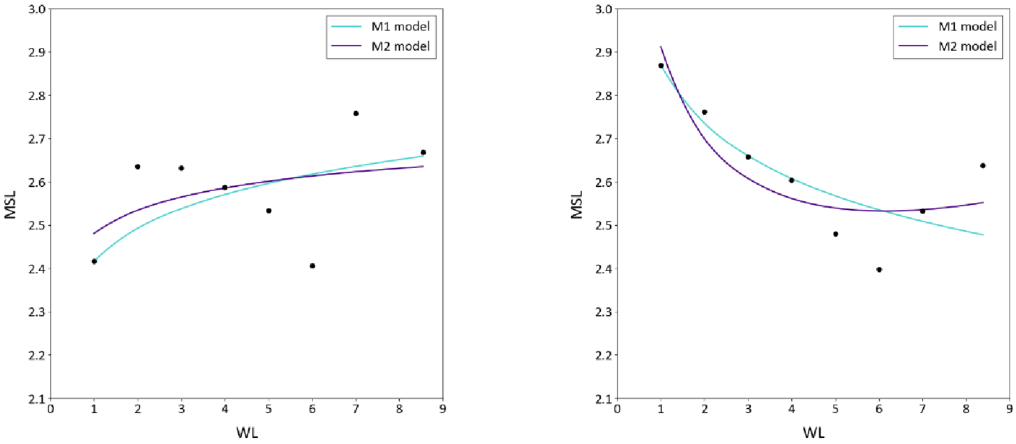
Figure 41. MAL applied to the triplet of the word, syllable and sound – tokens and types.

The hypothesis is rejected when testing the law on the word tokens – R^2 of M1 is not available or is unsatisfactory and R^2 of M2 relates to the increase in mean syllable lengths. MSL_2 has the highest values and mean syllable lengths in HK-P even peak with MSL_3. However, testing the word types from the UD samples brings completely different results compared to the component and stroke. MSL_1s are the highest and the mean syllable lengths continue decreasing with only a few exceptions (i.e. MSL_4 in GSD and $MSLs$ of the longest WL in PUD and PUD-W, highlighted in yellow in Table 55). The standard of $R_2 \geq 0.90$ is met in PUD-N and GSD (M1 and M2), and in PUD and PUD-W (M2). Hence, the hypothesis is not rejected in these cases. It is noteworthy that the fit of M1 in PUD is slightly below the standard ($R_2 = 0.8990$). Only PUD-W yields worse fitting results ($R_2 = 0.8645$). Regarding HK-P, we do not evaluate the results because its types have only three different word lengths after pooling the data, which are even fitted by three parameters in the case of M2 (hence, $R^2 > 0.99$).

As shown in the previous sub-chapters on the component and stroke, word segmentation seems to be again a decisive factor for the law to manifest itself and the menzerathian tendency to appear. To test whether the hypothesis's corroboration (with few exceptions) is UD specific when the sub-constituent is changed to the sound, we apply the law to LCMC, LCMC:M and LCMC:A+J. The results are shown in Table 56 and Figure 42 for LCMC and in Table 57 and Figure 43 for LCMC:M and LCMC:A+J.

Table 56. MAL applied to the triplet of the word, syllable and sound – tokens and types from LCMC.

LCMC: tokens			LCMC: types		
WL	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*
1	399331	2.42	1	322	2.87
2	390103	2.64	2	17882	2.76
3	24885	2.63	3	5922	2.66
4	12730	2.59	4	4357	2.60
5	614	2.53	5	224	2.48
6	133	2.41	6	88	2.40
7	125	2.76	7	37	2.53
8.54	24	2.67	8.39	18	2.64
	M1	M2		M1	M2
a	2.42	2.48	*a*	2.87	2.84
b	0.04	0.03	*b*	-0.07	-0.14
c		0.001	*c*		-0.02
R^2	0.1071	0.1796	R^2	0.6667	0.7580



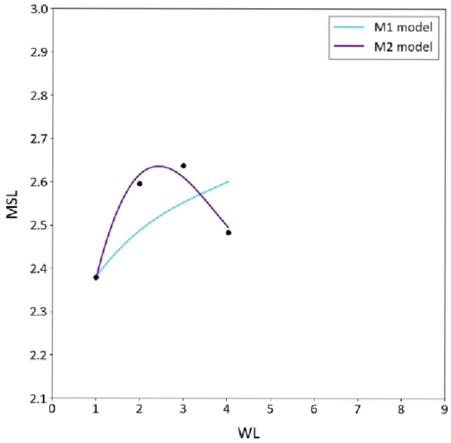
LCMC – tokens

LCMC – types

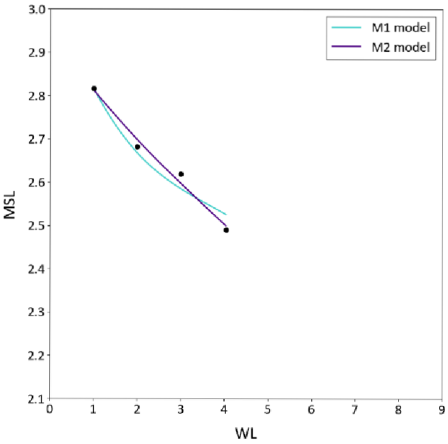
Figure 42. MAL applied to the triplet of the word, syllable and sound – tokens and types from LCMC.

Table 57. MAL applied to the triplet of the word, syllable and sound – tokens and types from LCMC:M and LCMC:A+J.

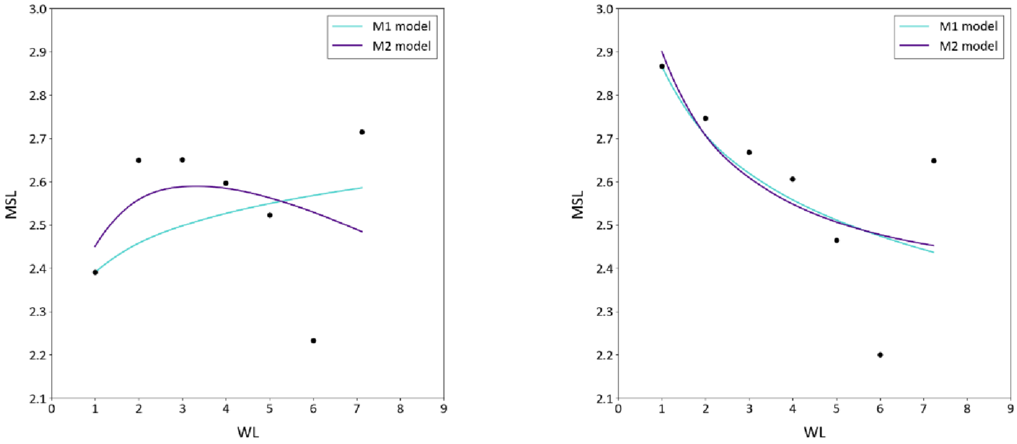
LCMC:M tokens			LCMC:M types			LCMC:A+J tokens			LCMC:A+J types		
WL	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*	*WL*	*f(WL)*	*MSL*
1	5064	2.38	1	223	2.82	1	86475	2.39	1	314	2.87
2	4661	2.60	2	1885	2.68	2	108331	2.65	2	9874	2.75
3	241	2.64	3	139	2.62	3	7407	2.65	3	2422	2.67
4.03	89	2.48	4.04	81	2.49	4	3265	2.60	4	1418	2.61
						5	161	2.52	5	74	2.46
						6	30	2.23	6	20	2.20
						7.12	25	2.71	7.23	13	2.65
	M1	M2		M1	M2		M1	M2		M1	M2
a	2.38	2.74	*a*	2.82	2.91	*a*	2.39	2.53	*a*	2.87	2.88
b	0.06	0.35	*b*	-0.08	-0.01	*b*	0.04	0.11	*b*	-0.08	-0.11
c		0.14	*c*		0.03	*c*		0.03	*c*		-0.01
R^2	0.1890	0.9662	R^2	0.9531	0.9853	R^2	NA	0.0926	R^2	0.5374	0.5432



LCMC:M – tokens



LCMC:M – types



LCMC:A+J – tokens

LCMC:A+J – types

Figure 43. MAL applied to the triplet of the word, syllable and sound – tokens and types from LCMC:M and LCMC:A+J.

The word tokens do not corroborate the hypothesis as in the UD case. MSL_1 has lower values than MSL_2 in all three samples and R^2 of both the models is unsatisfactory. Contrary to UD, the standard of $R^2 \geq 0.90$ is neither met by the word types except for LCMC:M. Despite the poor goodness-of-fit yielded by LCMC and LCMC:A+J (e.g. $R^2 = 0.6677$ and $R^2 = 0.5374$ of M1 accordingly), MSL_1 reaches the highest values contrary to the tokens, and mean syllable lengths continue decreasing except for the last $MSLs$. Only the LCMC:M types corroborate the hypothesis. As mentioned in Chapter 4.4.1, LCMC:M does not include words having seven and more syllables which violate the decreasing tendency in LCMC and LCMC:A+J. If we exclude these words, the fit of M1 improves, i.e. $R^2 = 0.9163$ while $b = -08$ in LCMC, and $R^2 = 0.7580$ while $b = -0.10$ in LCMC:A+J.

Despite the various results obtained when testing this triplet on different samples, word segmentation and the analysis of unit types represent crucial factors for the law to manifest itself and the menzerathian tendency to appear. If we evaluate the word lengths and the mean syllable lengths by the optics of the short-term memory span, we can conclude that they meet the upper threshold without exception (i.e. 7±2, Miller, 1956).

4.4.4 Summary of triplets on the word level

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 (see Table 58). 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 when it comes to a sub-constituent (cf. the sentence level, Chapter 4.1.5). 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.

Table 58. The parameters (a, b, c) and the coefficient of determination R^2 of both the models (M1, M2) obtained on the word level.

	HK-P		PUD		PUD-N		PUD-W		GSD		LCMC		LCMC:M		LCMC:A+J	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
word-character-component – tokens																
a	2.07	2.87	1.87	2.06	1.91	2.11	1.83	2.01	1.93	2.03	2.04	2.11	2.03	2.27	2.00	2.20
b	-0.01	0.67	0.14	0.34	0.14	0.35	0.15	0.33	0.11	0.19	0.01	0.04	0.04	0.26	0.02	0.27
c		0.34		0.09		0.10		0.08		0.03		0.01		0.11		0.10
R^2	NA	0.9055	0.4643	0.9567	0.6465	0.9749	0.5677	0.9414	0.5571	0.7748	NA	0.1158	0.0368	0.9990	NA	0.9647
word-character-component – types																
a	2.19	2.43	2.26	2.31	2.22	2.29	2.20	2.24	2.47	2.40	2.67	2.58	2.35	2.47	2.56	2.58
b	-0.01	0.18	0.01	0.07	0.03	0.11	0.03	0.07	-0.04	-0.08	-0.14	-0.24	-0.08	0.02	-0.14	-0.09
c		0.11		0.02		0.03		0.02		-0.02		-0.03		0.05		0.02
R^2	0.1964	>0.99	NA	0.4341	0.4632	0.9278	0.4946	0.6648	0.0039	0.3960	0.8866	0.9771	0.9343	0.9987	0.9422	0.9519
word-character-component – tokens excluding proper nouns																
a	2.07	2.82	1.87	2.37	1.90	2.27	1.83	2.54	1.91	2.06	2.03	2.12	2.01	2.27	1.99	2.23
b	-0.02	0.61	0.14	0.61	0.14	0.48	0.13	0.79	0.07	0.15	0.0001	0.04	0.05	0.29	0.00	0.26
c		0.32		0.24		0.17		0.33		0.05		0.02		0.12		0.10
R^2	0.0337	0.9696	0.3342	>0.99	0.5684	0.9894	0.0477	0.9856	0.0031	0.1463	NA	0.1509	0.0901	0.9982	NA	0.7693
word-character-component – types excluded proper nouns																
a	2.17	2.64	2.24	2.50	2.20	2.36	2.18	2.68	2.42	2.33	2.66	2.63	2.35	2.50	2.55	2.63
b	-0.04	0.31	-0.01	0.21	0.03	0.16	-0.02	0.39	-0.08	-0.19	-0.17	-0.24	-0.08	0.04	-0.16	-0.10
c		0.19		0.11		0.07		0.21		-0.05		-0.02		0.06		0.03
R^2	0.3620	>0.99	0.0531	0.9442	0.1458	0.9976	0.0803	0.9307	0.6955	0.7885	0.8743	0.8895	0.9071	0.9997	0.8609	0.8799

	word-character-component – tokens, CHISE decomposition															
a	1.74	2.17	1.72	1.78	1.74	1.81	1.71	1.77	1.72	1.74	1.78	1.83	1.78	1.95	1.80	1.90
b	-0.003	0.43	0.04	0.10	0.05	0.14	0.04	0.11	0.03	0.01	-0.03	-0.003	-0.003	0.17	-0.03	0.12
c		0.22		0.03		0.04		0.03		0.00		0.01		0.09		0.06
R^2	NA	>0.99	0.3369	0.7638	0.5543	0.9806	0.2022	0.7670	0.4638	0.5213	0.5792	0.7212	0.0144	0.9933	0.4207	0.9427
	word-character-component – types, CHISE decomposition															
a	1.84	2.05	1.84	1.85	1.83	1.87	1.82	1.84	1.92	1.88	1.96	1.96	1.88	1.97	1.94	1.99
b	-0.04	0.15	0.001	0.01	0.01	0.06	0.004	0.03	-0.03	-0.10	-0.09	-0.11	-0.05	0.04	-0.09	-0.02
c		0.11		0.004		0.02		0.01		-0.02		-0.01		0.05		0.03
R^2	0.6563	>0.99	NA	0.3273	0.1045	0.5502	NA	0.5656	0.2267	0.9468	0.9316	0.9371	0.8876	0.9989	0.8844	0.9475
	word-character-component – tokens, BLCU maximal decomposition															
a	2.20	2.92	1.98	2.21	2.02	2.29	1.94	2.16	2.05	2.18	2.19	2.31	2.15	2.50	2.14	2.41
b	0.07	0.66	0.17	0.41	0.19	0.45	0.18	0.40	0.14	0.22	0.02	0.07	0.08	0.37	0.03	0.33
c		0.29		0.10		0.12		0.09		0.04		0.03		0.15		0.12
R^2	NA	0.9426	0.5594	0.9715	0.6787	0.9789	0.6121	0.9631	0.6377	0.8352	NA	0.2169	0.3089	0.9974	NA	0.9470
	word-character-component – types, BLCU maximal decomposition															
a	2.41	2.60	2.48	2.53	2.43	2.51	2.39	2.44	2.81	2.69	3.05	2.94	2.63	2.76	2.91	2.94
b	0.01	0.14	0.03	0.08	0.05	0.12	0.05	0.12	-0.05	-0.13	-0.17	-0.28	-0.09	0.001	-0.16	-0.11
c		0.07		0.02		0.03		0.03		-0.03		-0.04		0.05		0.02
R^2	NA	>0.99	0.2312	0.4809	0.6129	0.8875	0.4993	0.6426	NA	0.6275	0.9118	0.9812	0.9520	0.9993	0.9484	0.9575
	word-character-component – tokens, CHISE maximal decomposition															
a	2.07	2.89	2.04	2.14	2.07	2.23	2.01	2.13	2.04	2.08	2.15	2.21	2.12	2.32	2.17	2.32
b	-0.02	0.66	0.06	0.15	0.05	0.22	0.07	0.21	0.05	0.06	-0.05	-0.02	-0.003	0.18	-0.06	0.13
c		0.34		0.04		0.07		0.06		0.01		0.01		0.09		0.07
R^2	0.0560	0.9085	0.3159	0.7845	0.1609	0.9434	0.2009	0.9555	0.3938	0.4489	0.6700	0.7588	0.0127	0.9014	0.5554	0.9761

	word-character-component – types, CHISE maximal decomposition															
a	2.19	2.42	2.24	2.28	2.21	2.34	2.16	2.25	2.51	2.41	2.71	2.64	2.37	2.49	2.59	2.62
b	-0.03	0.15	0.002	0.04	0.01	0.14	0.02	0.11	-0.07	-0.16	-0.18	-0.25	-0.10	0.003	-0.17	-0.12
c		0.10		0.02		0.06		0.04		-0.03		-0.02		0.05		0.02
R^2	0.5384	>0.99	NA	0.4891	NA	0.8360	NA	0.8380	0.4858	0.8590	0.9472	0.9714	0.9511	0.9990	0.9515	0.9595
	word-character-stroke – tokens															
a	6.84	8.87	6.31	6.86	6.42	7.03	6.20	6.76	6.54	6.82	6.97	7.44	6.81	7.84	6.92	7.82
b	0.02	0.55	0.15	0.33	0.16	0.36	0.16	0.35	0.12	0.18	-0.03	0.07	0.07	0.35	-0.02	0.33
c		0.27		0.08		0.09		0.08		0.02		0.04		0.14		0.13
R^2	NA	0.9474	0.6882	0.9909	0.7776	0.9935	0.6859	0.9881	0.7166	0.8279	0.2960	0.6671	0.2378	0.9969	0.0684	0.9373
	word-character-stroke – types															
a	7.24	8.06	7.91	7.90	7.70	7.80	7.41	7.56	8.88	8.46	9.80	9.64	8.28	8.69	9.32	9.61
b	0.02	0.21	0.01	0.01	0.03	0.06	0.05	0.10	-0.06	-0.16	-0.22	-0.27	-0.09	0.001	-0.20	-0.08
c		0.11		0.001		0.01		0.02		-0.04		-0.02		0.05		0.04
R^2	NA	>0.99	0.1071	0.1137	0.6278	0.7197	0.4882	0.5786	NA	0.7100	0.9387	0.9457	0.9525	0.9990	0.9132	0.9428
	word-syllable-sound – tokens															
a	2.39	2.85	2.34	2.61	2.35	2.75	2.33	2.60	2.44	2.59	2.42	2.48	2.38	2.74	2.39	2.53
b	0.06	0.44	-0.01	0.18	0.005	0.34	-0.02	0.17	-0.04	0.04	0.04	0.03	0.06	0.35	0.04	0.11
c		0.18		0.09		0.15		0.09		0.04		0.001		0.14		0.03
R^2	0.0193	0.7319	0.0715	0.6113	NA	0.9383	0.0987	0.5955	0.4320	0.6935	0.1071	0.1796	0.1890	0.9662	NA	0.0926
	word-syllable-sound – types															
a	2.76	2.53	2.81	2.90	2.79	2.99	2.75	2.87	2.81	2.85	2.87	2.84	2.82	2.91	2.87	2.88
b	-0.04	-0.19	-0.14	-0.09	-0.12	0.03	-0.13	-0.06	-0.14	-0.15	-0.07	-0.14	-0.08	-0.01	-0.08	-0.11
c		-0.08		0.02		0.07		0.03		-0.003		-0.02		0.03		-0.01
R^2	0.5469	>0.99	0.8990	0.9218	0.9038	0.9997	0.8645	0.9024	0.9102	0.9154	0.6667	0.7580	0.9531	0.9853	0.5374	0.5432

4.5 The character as the construct

4.5.1 The component and stroke as constituents

Hypothesis: the longer the Chinese character length measured in the number of components, the shorter the mean length of the components measured in strokes.[146]

The results of character tokens and types decomposed while using the BLCU source are presented in Table 59 and Table 60 accordingly and in

Figure 44. ChL stands for the character length measured the in the number of components, $f(ChL)$ for its frequency and $MCoL$ for the mean component length measured in the number of strokes. Both the models are applied to the data – the truncated model $y(x) = ax^b$ with the M1 label and the complete model $y(x) = ax^b e^{cx}$ with the M2 label. The parameters (a, b, c) and the coefficient of determination R^2 are presented in the tables. When fitting the data with M1, we use the mean component length of one-component characters, i.e. $MCoL_1$, as the parameter a.

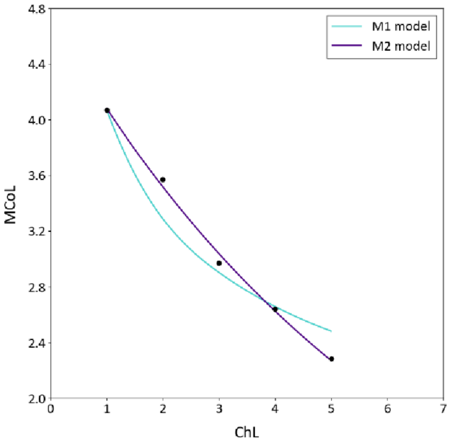
[146] All non-Chinese graphemes are excluded from the analysis.

Table 59. MAL applied to the triplet of the character, component and stroke – tokens.

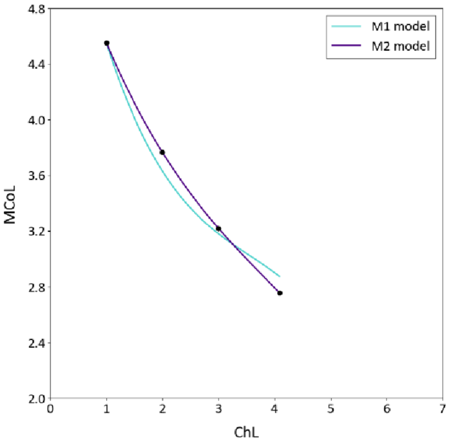
HK-P			PUD			PUD-N			PUD-W			GSD		
ChL	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*
1	1013	4.07	1	6014	3.95	1	2754	3.95	1	3260	3.95	1	26919	4.09
2	3790	3.57	2	15262	3.58	2	7378	3.60	2	7884	3.56	2	64568	3.59
3	1613	2.97	3	6625	3.06	3	3103	3.05	3	3522	3.08	3	29952	3.09
4	362	2.64	4	1544	2.84	4	777	2.85	4	767	2.84	4	6439	2.88
5	12	2.28	5	315	2.79	5.05	161	2.81	5.02	164	2.77	5	1320	2.73
			6.10	10	2.75							6.14	50	2.68
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	4.07	4.71	*a*	3.95	3.98	*a*	3.95	4.10	*a*	3.95	4.12	*a*	4.09	4.13
b	-0.31	-0.01	*b*	-0.21	-0.23	*b*	-0.22	-0.16	*b*	-0.22	-0.15	*b*	-0.24	-0.24
c		0.14	*c*		-0.003	*c*		0.03	*c*		0.04	*c*		0.003
R^2	0.9398	0.9962	R^2	0.9619	0.9636	R^2	0.9504	0.9591	R^2	0.9678	0.9785	R^2	0.9829	0.9841

Table 60. MAL applied to the triplet of the character, component and stroke – types.

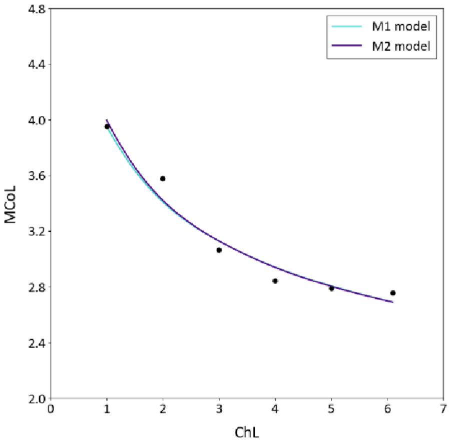
HK-P			PUD			PUD-N			PUD-W			GSD		
ChL	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*
1	76	4.55	1	172	4.51	1	147	4.44	1	165	4.52	1	210	4.65
2	305	3.77	2	910	4.03	2	734	3.98	2	746	3.93	2	1425	4.20
3	175	3.22	3	667	3.39	3	504	3.31	3	519	3.34	3	1160	3.54
4.09	44	2.76	4	204	3.13	4	156	3.12	4	146	3.09	4	393	3.19
			5.17	41	2.97	5.13	32	2.94	5.11	28	2.95	5	79	3.01
												6.10	21	2.91
	M1	M2		M1	M2		M1	M2		M1	M2		M1	M2
a	4.55	5.08	*a*	4.51	4.77	*a*	4.44	4.68	*a*	4.52	4.66	*a*	4.65	4.91
b	-0.33	-0.12	*b*	-0.25	-0.15	*b*	-0.24	-0.15	*b*	-0.26	-0.21	*b*	-0.25	-0.16
c		0.11	*c*		0.05	*c*		0.05	*c*		0.03	*c*		0.04
R^2	0.9812	>0.99	R^2	0.9614	0.9768	R^2	0.9579	0.9719	R^2	0.9813	0.9857	R^2	0.9564	0.9746



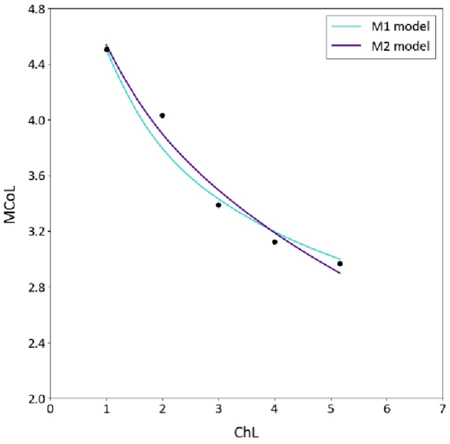
HK-P – tokens



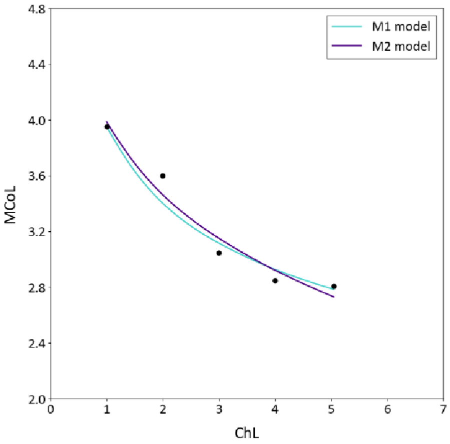
HK-P – types



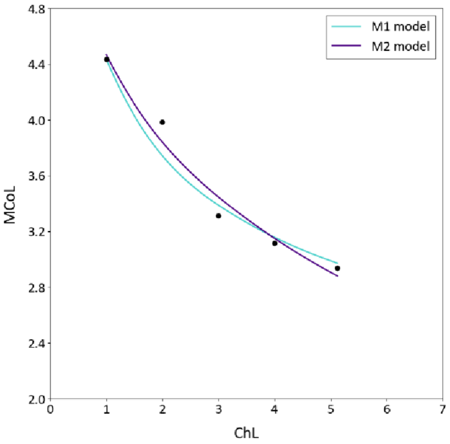
PUD – tokens



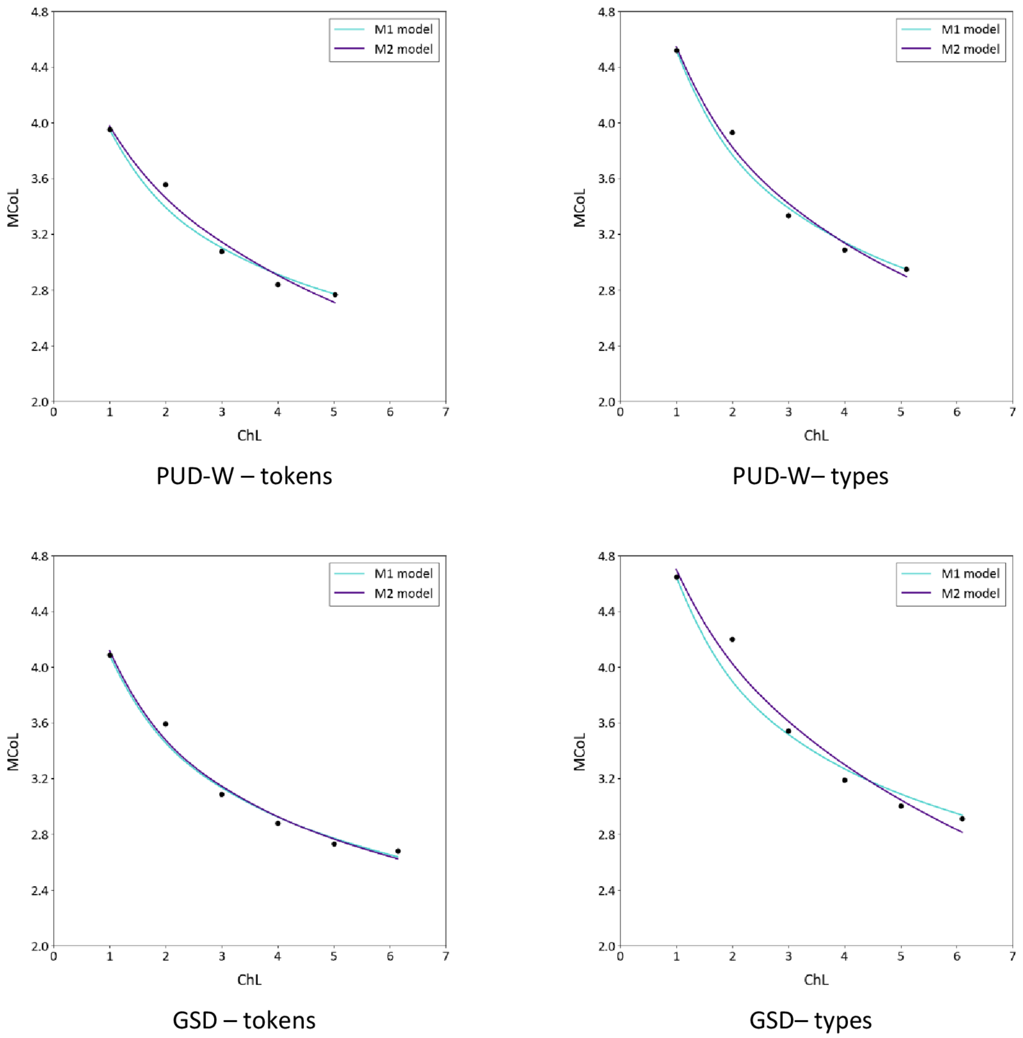
PUD– types



PUD-N – tokens



PUD-N– types



PUD-W – tokens

PUD-W– types

GSD – tokens

GSD– types

Figure 44. MAL applied to the triplet of the character, component and stroke – tokens and types.

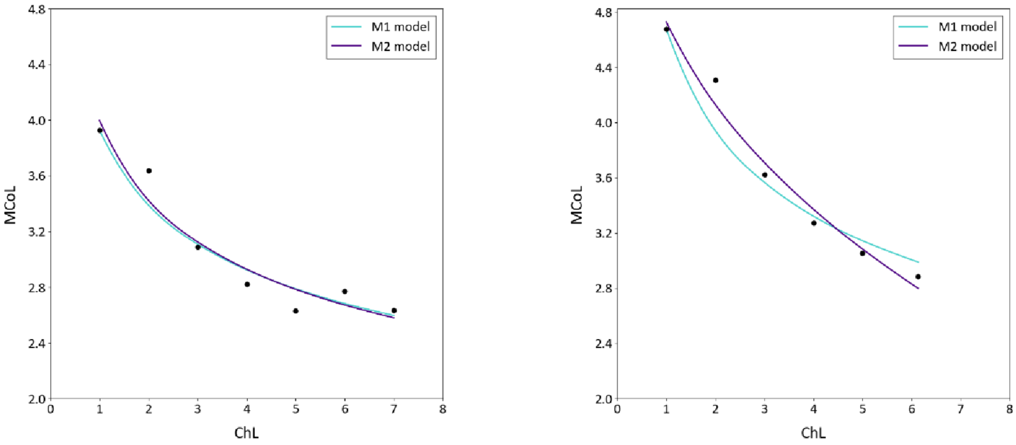
The triplet consisting of the character, component and stroke is the only unit combination which corroborates the hypothesis for both – the tokens and the types. The goodness-of-fit between the models and all the data meets the standard of $R^2 \geq 0.90$ and the menzerathian decreasing tendency of $MCoL$ is not even violated by the second regime.

As mentioned above, the tokens do not reject the hypothesis contrary to the findings on higher linguistic levels. The question arises whether they also behave in accord with the law in a large sample such as LCMC.[147] The results of LCMC character tokens and types, which we decompose using the BLCU source, are presented in Table 61 and Figure 45.

[147] We remind the reader that LCMC contains more than 800k word tokens and GSD – the largest UD treebank – slightly below 81k (see Chapter 3.1).

Table 61. MAL applied to the triplet of the character, component and stroke – tokens and types from LCMC.

LCMC: tokens			LCMC: types		
ChL	*f(ChL)*	*MCoL*	*ChL*	*f(ChL)*	*MCoL*
1	276911	3.93	1	232	4.68
2	674218	3.64	2	1840	4.31
3	283098	3.09	3	1773	3.62
4	67898	2.82	4	664	3.27
5	10948	2.63	5	143	3.05
6	736	2.77	6.14	37	2.89
7	137	2.63			
	M1	M2		M1	M2
a	3.93	4.00	*a*	4.68	5.06
b	-0.21	-0.22	*b*	-0.25	-0.10
c		0.001	*c*		0.07
R^2	0.9326	0.9369	R^2	0.9381	0.9766



LCMC – tokens

LCMC – types

Figure 45. MAL applied to the triplet of the character, component and stroke – tokens and types from LCMC.

The LCMC sample corroborates the hypothesis not only for the types but also for the tokens. Contrary to the UD samples, the tokens from LCMC show an increase in $MCoL_6$ (highlighted in yellow in Table 61). Despite the second regime, the goodness-of-fit meets the standard of $R^2 \geq 0.90$ as well as in case of the types where the second regime does not occur.

We tested on the word level whether different approaches to the decomposition of Chinese characters have an impact on the results. Firstly, we decomposed Chinese characters based on an alternative source called CHISE. Secondly, we maximally decomposed each character into its components until all identified components could not be decomposed further while using the BLCU and CHISE source (for more information, see Chapter 4.4.1). The results showed that the law might be sensitive to a decomposing approach. The question arises whether this also applies when the Chinese character switches its position over to the construct. Hence, we also quantify the size of the Chinese characters while using alternative decompositions. However, due to the fact that the CHISE decomposition mostly results in three character lengths when the types are analysed, we present only the results yielded by the maximal decompositions, which do not suffer from this drawback. Table 62 shows the results of the tokens and the types from all the samples (including LCMC).

Table 62. The parameters (a, b, c) and the coefficients of determination R^2 of both the models (M1, M2) obtained by the analysis of tokens and types of Chinese characters based on the maximal BLCU and CHISE decomposition – UD samples and LCMC.

	HK-P		PUD		PUD-N		PUD-W		GSD		LCMC	
	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2	M1	M2
	(1) tokens - BLCU maximal decomposition											
a	4.07	4.04	3.95	4.01	3.95	3.98	3.95	4.13	4.09	4.17	3.93	4.10
b	-0.32	-0.35	-0.29	-0.29	-0.28	-0.31	-0.30	-0.21	-0.32	-0.27	-0.30	-0.19
c		-0.01		0.003		-0.01		0.04		0.02		0.04
R^2	0.9860	0.9866	0.9731	0.9745	0.9759	0.9772	0.9798	0.9936	0.9804	0.9853	0.9350	0.9636
	(2) tokens - CHISE maximal decomposition											
a	4.73	3.57	4.17	4.49	4.18	4.41	4.15	4.49	4.28	4.71	4.15	4.50
b	-0.34	-0.98	-0.31	-0.15	-0.29	-0.20	-0.31	-0.14	-0.33	-0.12	-0.30	-0.13
c		-0.30		0.07		0.05		0.07		0.09		0.08
R^2	0.6701	0.8494	0.9651	0.9896	0.9761	0.9853	0.9643	0.9907	0.9629	0.9966	0.9631	0.9927
	(3) types - BLCU maximal decomposition											
a	4.55	4.53	4.51	4.74	4.44	4.64	4.52	4.68	4.65	4.87	4.68	4.94
b	-0.38	-0.39	-0.34	-0.22	-0.33	-0.23	-0.34	-0.26	-0.34	-0.24	-0.33	-0.21
c		-0.004		0.05		0.04		0.03		0.04		0.05
R^2	0.9998	0.9999	0.9831	0.9983	0.9882	0.9977	0.9930	0.9982	0.9800	0.9958	0.9743	0.9964
	(4) types - CHISE maximal decomposition											
a	4.67	5.22	4.81	5.30	4.78	5.22	4.82	5.26	5.03	5.44	5.11	5.56
b	-0.36	-0.15	-0.34	-0.13	-0.34	-0.16	-0.35	-0.18	-0.35	-0.19	-0.35	-0.16
c		0.11		0.10		0.09		0.09		0.07		0.08
R^2	0.9831	0.9993	0.9752	0.9991	0.9796	0.9987	0.9820	0.9993	0.9770	0.9975	0.9739	0.9979

The hypothesis is rejected only for HK-P when both the models are applied to the tokens. Otherwise, the goodness-of-fit reaches the standard of $R^2 \geq 0.90$ without regard to the tokens or the types and the source used for the maximal decomposition. Hence, neither the alternative decompositions bring considerable changes to the results.

Let us evaluate the values of the M1 parameters yielded in all the samples if $R^2 \geq 0.90$ (Figure 46). The parameter a, i.e. $MCoL_1$, tends to be under the influence of the unit frequency – analysing the character types results in its higher values compared to the character tokens. On the other hand, values of the parameter b mainly change with the decomposing approach to the characters. While the character tokens and types which are decomposed using the BLCU source yield higher values of the parameter b, the maximal BLCU decomposition lowers its values. As for the relationship between both the parameters, the negative correlation of their values is indicated.

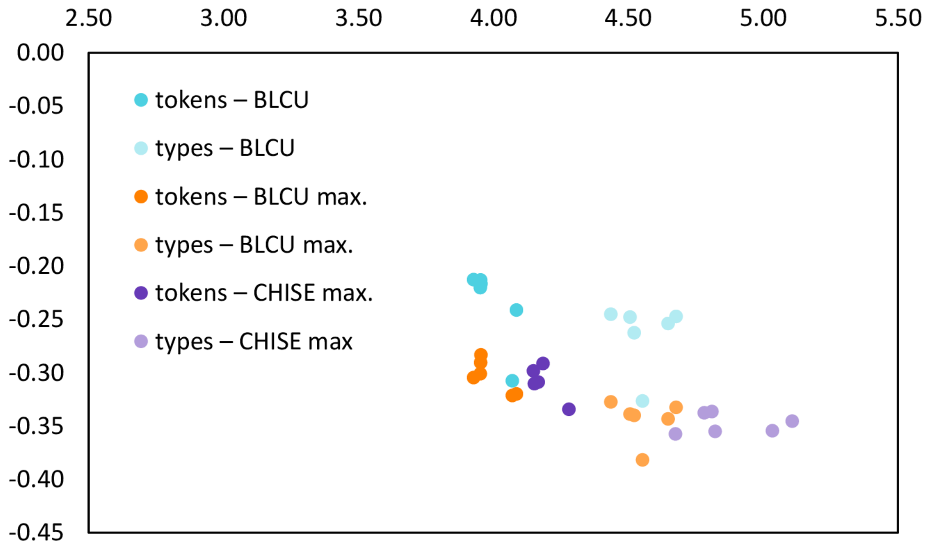


Figure 46. The parameters a and b of M1 for the triplet of the character, component and stroke.

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.

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

As regards the interpretation of parameters, most attention is focused on the parameters a and b of the truncated model, i.e. $y = ax^b$, rather than on parameters of the complete formula, i.e. $y = ax^b e^{-cx}$, due to easier linguistic interpretability. Both the parameters are expected to be influenced by a linguistic level under analysis (e.g. Cramer, 2005b; Köhler 2012), but other factors influencing their values have been addressed too, e.g. an influence of a text type (Teupenhayn and Altmann, 1984; Kułacka, 2010; Čech and Mačutek, 2021). Moreover, values of the parameters appear to be negatively correlated – with an increasing value of the parameter a, the value of the parameter b decreases (e.g. Hammerl and Sambor, 1993; Hou et al., 2019a; Jiang and Jiang, 2022).

This last chapter on the results presents an overview 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. The overview is presented in Table 63 and Figure 47 and includes only those values which we obtained when the coefficient of determination R^2 met the standard, i.e. $R^2 \geq 0.90$. 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, as discussed within the previous chapters.

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 shifts the values on the x axis to the right, or in other words, 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 type (cf. HK-P vs PUD and GSD).

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 (see Figure 47). If we apply the Kendall rank correlation test to all values of the parameters from Table 63 (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).

However, several issues arise when evaluating the values of the parameters and their relationship. Firstly, we face the choice of a model and its impact on the results. We used the truncated formula, but studies also opt for the complete model. Secondly, we analysed only values obtained when $R^2 \geq 0.90$, which poses the question of what is the minimum threshold of R^2 for the parameters to be evaluated, or even more general, for the law to be corroborated. Thirdly, a chosen methodology (e.g. determination of language units, a sample under analysis, homogeneity) influences the results, which puts comparability into question.

Table 63. Overview of the parameters a and b of the truncated model obtained from linguistic levels under analysis.

	HK-P		PUD		PUD-N		PUD-W		GSD		LCMC		LCMC:M		LCMC:A+J	
	a	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
									Sentence level							
sentence, clause, word	5.63	-0.25	12.50	-0.59	10.99	-0.52	13.87	-0.66	11.64	-0.48	not tested		not tested		not tested	
sentence, clause (punctuation), word	7.85	-0.27	12.63	-0.44	12.55	-0.47	12.73	-0.43	13.10	-0.39						
sentence, phrase, word			11.27	-0.66	13.16	-0.79	12.13	-0.70	19.03	-0.90						
sentence, clause, phrase	3.22	-0.33	3.99	-0.43	3.95	-0.43	4.03	-0.43	3.66	-0.37						
sentence, clause (0-phrase clauses excl.), phrase	3.17	-0.22	4.01	-0.38	3.97	-0.38	4.04	-0.39								
sentence, clause, LDS	3.49	-0.25	6.99	-0.55	6.39	-0.50	7.54	-0.61	6.46	-0.44						
									Clause level							
clause, word, character											not tested		not tested		not tested	
clause (punctuation), word, character																
clause, phrase, word																
clause (heads incl.), phrase, word									3.31	-0.28						
clause, LDS, word																

	HK-P		PUD		PUD-N		PUD-W		GSD		LCMC		LCMC:M		LCMC:A+J	
	a	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
Phrase level																
sentential phrase, word, character											not tested		not tested		not tested	
clausal phrase, word, character – tokens																
clausal phrase, word, character – types																
clausal phrase, word, character – Chinese tokens																
clausal phrase, word, character – Chinese types																
LDS, word, character – tokens																
LDS, word, character – Chinese types			2.13	-0.21	2.01	-0.19	2.13	-0.19								
Word level – types																
word, character, component – BLCU													2.35	-0.08	2.56	-0.14
word, character, component – BLCU, proper nouns excl.													2.35	-0.08		
word, character, component – CHISE											1.96	-0.09				
word, character, component – BLCU max.											3.05	-0.17	2.63	-0.09	2.91	-0.16
word, character, component – CHISE max.											2.71	-0.18	2.37	-0.10	2.59	-0.17
word, character, stroke											9.80	-0.22	8.28	-0.09	9.32	-0.20
word, syllable, sound					2.79	-0.12			2.81	-0.14			2.82	-0.08		

	HK-P		PUD		PUD-N		PUD-W		GSD		LCMC		LCMC:M		LCMC:A+J	
	a	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
	Character level															
character, component, stroke – tokens, BLCU	4.07	-0.31	3.95	-0.21	3.95	-0.22	3.95	-0.22	4.09	-0.24	3.93	-0.21	not tested		not tested	
character, component, stroke – types, BLCU	4.55	-0.33	4.51	-0.25	4.44	-0.24	4.52	-0.26	4.65	-0.25	4.68	-0.25				
character, component, stroke – tokens, BLCU max.	4.07	-0.32	3.95	-0.29	3.95	-0.28	3.95	-0.30	4.09	-0.32	3.93	-0.30				
character, component, stroke – types, BLCU max.	4.55	-0.38	4.51	-0.34	4.44	-0.33	4.52	-0.34	4.65	-0.34	4.68	-0.33				
character, component, stroke – tokens, CHISE max.			4.17	-0.31	4.18	-0.29	4.15	-0.31	4.28	-0.33	4.15	-0.30				
character, component, stroke – types, CHISE max.	4.67	-0.36	4.81	-0.34	4.78	-0.34	4.82	-0.35	5.03	-0.35	5.11	-0.35				

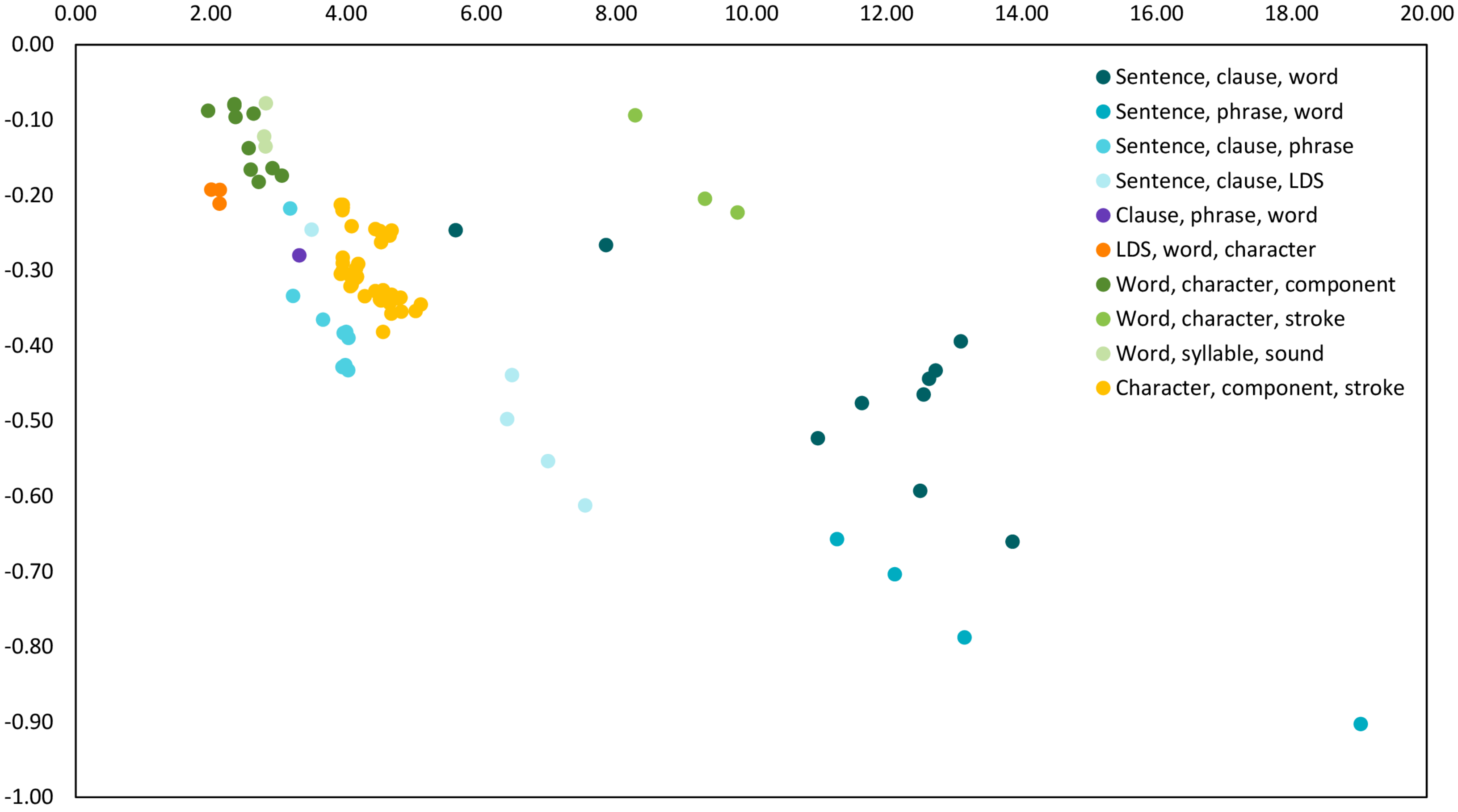


Figure 47. Visualisation of the parameters a and b of the truncated model obtained from all linguistic levels under analysis.

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., 2021b). 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.

Summary in the Czech language

Disertační práce se zaměřuje na Menzerath-Altmannův zákon, který je v rámci práce testován na čínském jazykovém materiálu. Podle tohoto zákona délky dvou jazykových jednotek různých hierarchických úrovní – hierarchicky vyššího konstruktu a hierarchicky nižšího konstituentu – spolu negativně korelují. Platnost zákona byla v posledních čtyřech desetiletích ověřena na různých jazycích, jazykových jednotkách a různém jazykovém materiálu. Určité jednotky však přitahují více pozornosti než jiné a obvykle se testuje pouze jedna dvojice jednotek (první v pozici konstruktu, druhá v pozici konstituentu), přestože jednotka může svoji hierarchickou pozici změnit. Zároveň se předpokládá, že se negativní korelace mezi délkami jednotek objevuje, pokud se analyzují bezprostředně sousedící jednotky. Hranice mezi jednotkami však nejsou vždy zřejmé. I když již dvě studie aplikovaly Menzerath-Altmannův zákon na hierarchii jazykových jednotek v čínštině (Chen a Liu, 2019, 2022), obě zvolily klauzi a slovo jako bezprostředně sousedící jednotky a rovinu fráze vynechaly z analýzy. Studie zároveň ukázaly, že zákon nevstupuje v platnost, pokud se aplikuje na slovo v pozici konstruktu a čínský znak v pozici jeho konstituentu. S ohledem na výše uvedené disertační práce aplikuje Menzerath-Altmannův zákon napříč hierarchií jazykových jednotek, která je složená z věty, klauze, fráze, slova, znaku/slabiky, komponentu/hlásky a tahu. Práce nejprve sleduje, jak se chovají neperiferní jednotky, když se změní jejich pozice z konstituentu na konstrukt. Zároveň je zákon aplikován na různé kombinace jednotek, včetně fráze v různých pozicích, z důvodu testování hranic mezi jednotkami v čínštině. V neposlední řádě práce zkoumá, zda existují faktory (např. frekvence), které brání zákonu projevit se na rovině slova. Na základě dosažených výsledků lze konstatovat několik závěrů. Roviny věty, fráze, slova a čínského znaku v pozici konstruktu jsou v souladu se zákonem, zatímco klauze přináší opačné výsledky. Pokud jde o hranice mezi jednotkami, klauze a slovo se nejeví jako bezprostředně sousedící jednotky stejně jako věta a fráze. Předběžně lze zároveň konstatovat, že fráze patří do hierarchie jazykových jednotek v čínštině a délka slova měřená v počtu čínských znaků nebrání zákonu vstoupit v platnost bez ohledu na hierarchickou pozici. Na druhou stranu výsledky také ukazují, že na zákon a jeho projevení se má vliv několik zásadních faktorů. Zaprvé, v případě frází a slov se zákon projevuje pouze tehdy, pokud se analyzují jejich tzv. typy (types) a nikoli tokeny (tokens), tj. nebere se v úvahu frekvence. Zadruhé, analýza fráze ukazuje, že zákon je citlivý na homogenitu jazykového materiálu, tj. vstupuje v platnost, pokud je aplikován na typy frází neobsahující nečínské znaky. Závěrem se zákon projevuje v závislosti na způsobu segmentace na slova (tzv. tokenizace), která je uplatněná v rámci analyzovaného jazykového materiálu. Při testování Menzerath-Altmannova zákona je tedy důležité zohlednit konkurující a kooperující faktory, které jednak mohou ovlivnit výsledky, jednak poodhalit chování jazykových jednotek i samotného zákona.

Bibliography

Acl: clausal modifier of noun (adnominal clause) (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/acl.html, accessed: 30 April 2022.

Acl: clausal modifier of noun (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/acl.html, accessed: 30 April 2022.

Advcl: adverbial clause modifier (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/advcl.html, accessed: 30 April 2022.

Advcl: adverbial clause modifier (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/advcl.html, accessed: 30 April 2022.

Alekseev, P. M. (1998) "Graphemic and Syllabic length of words in text and vocabulary", *Journal of Quantitative Linguistics*, 5(1-2), pp. 5-12, doi: 10.1080/09296179808590107.

Altmann, E. G. and Gerlach, M. (2016) "Statistical Laws in Linguistics", in Degli Esposti, M., Altmann, E. G. and Pachet, F. (eds.) *Creativity and Universality in Language: Lecture Notes in Morphogenesis*, Springer, Cham, pp. 7-26, doi: 10.1007/978-3-319-24403-7_2.

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. (1988) "Verteilungen der Satzlängen", in Schulz, K.-P. (ed.) *Glottometrika 9*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 147-169.

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.

Altmann, G., Beöthy, E. and Best, K.-H. (1982) "Die Bedeutungskomplexität der Wörter und das Menzerathsche Gesetz", *STUF – Language Typology and Universals*, 35(JG), pp. 537-543, doi: 10.1524/stuf.1982.35.jg.537.

Altmann, G. et al. (2002) *Einführung in die quantitative Lexikologie*, Peust & Gutschmidt, Göttingen.

Andres, J. (2010) "On a Conjecture about the Fractal Structure of Language", *Journal of Quantitative Linguistics*, 17(2), pp. 101-122, doi: 10.1080/09296171003643189.

Andres, J. (2014) "The Moran–Hutchinson formula in terms of Menzerath–Altmann's law and Zipf–Mandelbrot's law", in Altmann, G. et al. (eds.) *Empirical Approaches to Text and Language Analysis: dedicated to Luděk Hřebíček on the occasion of his 80th birthday*, RAM-Verlag, Lüdenscheid, pp. 29-44.

Andres, J. (2017) "Hypotéza fraktální struktury jazyka (The hypothesis of a fractal structure of a language)", in Karlík, P., Nekula, M. and Pleskalová, J. (eds.) *CzechEncy – Nový encyklopedický slovník češtiny (CzechEncy – A New Encyclopedic Dictionary of the Czech language)*, available at: https://www.czechency.org/slovnik/HYPOT%C3%89ZA%20FRAKT%C3%81LN%C3%8D%20STRUKTURY%20JAZYKA , accessed: 27 March 2022.

Andres, J. and Benešová, M. (2011) “Fractal analysis of Poe’s Raven”, *Glottometrics*, 21, pp. 73-98, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g21zeit.pdf, accessed: 30 April 2022.

Andres, J. and Benešová, M. (2012) “Fractal Analysis of Poe’s Raven, II*”, *Journal of Quantitative Linguistics*, 19(4), pp. 301-324, doi: 10.1080/09296174.2012.714538.

Andres, J. et al. (2012a) “Methodological Note on the Fractal Analysis of Texts*”, *Journal of Quantitative Linguistics*, 19(1), pp. 1-31, doi: 10.1080/09296174.2011.608604.

Andres, J. et al. (2012b) “Optimization of parameters in the Menzerath–Altmann law”, *Acta Universitatis Palackianae Olomucensis: Facultas Rerum Naturalium. Mathematica*, 51(1), pp. 5-27.

Araujo, L., Benevides, A. and Pereira, M. (2020) “Análise da Lei de Menzerath no Português Brasileiro”, *Linguamática*, 12(1), pp. 31-48, doi: 10.21814/lm.12.1.300.

Arens, H. (1965) *Verborgene Ordnung*, Schwann, Düsseldorf.

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.

Benešová, M. (2011) *Kvantitativní analýza textu se zvláštním zřetelem k analýze fraktální (Quantitative Analysis of Text with Special Respect to Fractal Analysis)*, Ph.D. Dissertation, available at: https://theses.cz/id/p19fdf/DISERTACNI_PRACE_BENESOVA.pdf, accessed: 30 April 2022.

Benešová, M. and Birjukov, D. (2015) “Application of the Menzerath-Altmann Law to Contemporary Written Japanese in the Short Story Style”, in Tuzzi, A., Benešová, M. and Mačutek, J. (eds.) *Recent Contributions to Quantitative Linguistics*, De Gruyter Mouton, Berlin, München, Boston, pp. 13-25, doi: 10.1515/9783110420296-003.

Benešová, M. and Čech, R. (2015) “Menzerath-Altmann law versus random model”, in Mikros, G. K. and Mačutek, J. (eds.) *Sequences in Language and Text*, De Gruyter Mouton, Berlin, München, Boston, pp. 57-69, doi: 10.1515/9783110362879-005.

Benešová, M., Faltýnek, D. and Zámečník, L. H. (2015) “Menzerath-Altmann Law in Differently Segmented Texts”, in Tuzzi, A., Benešová, M. and Mačutek, J. (eds.) *Recent Contributions to Quantitative Linguistics*, De Gruyter Mouton, Berlin, München, Boston, pp. 35-48, doi: 10.1515/9783110420296-004.

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.

Best, K.-H. (2007) “XXIX. Paul Menzerath (1883-1954)”, *Glottometrics*, 14, pp. 86-98, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g14zeit.pdf, accessed: 30 April 2022.

Best, K.-H. and Rottmann, O. (2017) *Quantitative Linguistics: an Invitation*, RAM-Verlag, Lüdenscheid.

Birjukov, D. (2016) “Application of the Menzerath-Altmann Law to Written Japanese – Poetic and Academic Styles”, in Benešová, M. (ed.) *Text segmentation for Menzerath-Altmann law testing*, Palacký University Olomouc, Olomouc, pp. 7-43.

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. (2007) “Statistical parameters of Ivan Franko’s novel Perekhresni stežky (The Cross-Paths)”, 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. 39-48, doi: 10.1515/9783110894219.39.

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.

Carnegie Mellon University (n.d.) *The CMU Pronouncing Dictionary (version 0.7b)*, available at: http://www.speech.cs.cmu.edu/cgi-bin/cmudict#phones, accessed: 30 April 2022.

Ccomp: clausal complement (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/ccomp.html, accessed: 30 April 2022.

Ccomp: clausal complement (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/ccomp.html, accessed: 30 April 2022.

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.

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.

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.

CHISE: CHaracter Information Service Environment (2021), available at: https://www.chise.org/index.en.html, accessed: 17 June 2022).

Clink, D. J., Ahmad, A. H. and Klinck, H. (2020) “Brevity is not a universal in animal communication: evidence for compression depends on the unit of analysis in small ape vocalizations”, *Royal Society Open Science*, 7(200151), doi: 10.1098/rsos.200151.

Clink, D. J. and Lau, A. R. (2020) “Adherence to Menzerath’s Law is the exception (not the rule) in three duetting primate species”, *Royal Society Open Science*, 7(201557), doi: 10.1098/rsos.201557.

Coloma, G. (2015) “The Menzerath-Altmann Law in a Cross-Linguistic Context”, *SKY Journal of Linguistics*, 28, pp. 139-159, available at: http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Coloma.pdf, accessed: 30 April 2022.

Coloma, G. (2020) “Language Complexity Trade-Offs Revisited”, *Serie Documentos de Trabajo. Buenos Aires: Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA)*, 721, pp. 1-27, available at: https://www.econstor.eu/handle/10419/238346, accessed: 30 April 2022.

Complex Clauses (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/complex-syntax.html, accessed: 30 April 2022.

CoNLL-U Viewer, available at: https://universaldependencies.org/conllu_viewer.html, accessed: 30 April 2022.

Coordination (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/complex-syntax.html#coordination, accessed: 30 April 2022.

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.

Csubj: clausal subject (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/csubj.html, accessed: 30 April 2022.

Csubj: clausal subject (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/csubj.html, accessed: 30 April 2022.

Dekking, F. M. et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding Why and How*, Springer, London.

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

Dependencies (2021), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/index.html, accessed: 30 April 2022.

Dinu, A. and Dinu, L. P. (2009) “On the behavior of Romanian syllables related to minimum effort laws”, in Paskaleva, E. et al. (eds.) *Proceedings of the Workshop Multilingual resources, technologies and evaluation for central and Eastern {E}uropean languages*, Association for Computational Linguistics, Borovets, pp. 9-13, doi: 10.5555/1859119.1859121.

Favaro, L. et al. (2020) “Do penguins’ vocal sequences conform to linguistic laws?”, *Biology Letters*, 16(20190589), doi: 10.1098/rsbl.2019.0589.

Fedurek, P., Zuberbühler, K. and Semple, S. (2017) “Trade-offs in the production of animal vocal sequences: insights from the structure of wild chimpanzee pant hoots”, *Frontiers in Zoology*, 14(50), doi: 10.1186/s12983-017-0235-8.

Fenk, A. and Fenk-Oczlon, G. (1993) “Menzerath’s Law and the Constant Flow of Linguistic Information”, 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. 11-31.

Fenk, A., Fenk-Oczlon, G. and Fenk, L. (2005) “Syllable complexity as a function of word complexity”, in Solovyev, V. and Polyakov, V. (eds.) *Text Processing and Cognitive Technologies*, MISA, Moscow, pp. 324-333.

Fenk-Oczlon, G. and Fenk, A. (1995) “Selbstorganisation und natürliche Typologie”, *Sprachtypologie und Universalienforschung*, 48(3), pp. 223-238, doi: 10.1524/stuf.1995.48.3.223.

Ferrer-I-Cancho, R. and Forns, N. (2010) “The self-organization of genomes”, *Complexity*, 15, pp. 34-36, doi: 10.1002/cplx.20296.

Ferrer-i-Cancho, R. et al. (2014) “When is Menzerath-Altmann law mathematically trivial? A new approach”, *Statistical Applications in Genetics and Molecular Biology*, 13(6), pp. 633-644, doi: 10.1515/sagmb-2013-0034.

Fickermann, I., Markner-Jäger, B. and Rothe, U. (1984) “Wortlänge und Bedeutungskomplexität”, in Boy, J. and Köhler, R. (eds.) *Glottometrika 6*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 115-126.

Gajić, D. M. (1950) *Zur Struktur des serbokroatischen Wortschatzes: Die Typologie der serbokroatischen mehrsilbigen Wörter*, Ph.D. Dissertation.

Galieva, A. M. (2021) “Закон Мензерата–Альтманна: эксперименты с текстами на татарском языке (The Menzerath–Altmann Law: Experimenting with Tatar Texts)”, *Uchenye Zapiski Kazanskogo Universiteta. Seriya Gumanitarnye Nauki (Proceedings of Kazan University. Humanities Series)*, 163(1), pp. 180-189, doi: 10.26907/2541-7738.2021.1.180-189.

Gerdes, K. et al. (2018) “SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD”, in de Marneffe, M.-C., Lynn, T. and Schuster, S. (eds.) *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, Association for Computational Linguistics, Brussels, pp. 1-9, doi: 10.18653/v1/W18-6008.

Gerlach, R. (1982) “Zur Überprüfung des Menzerath’schen Gesetzes im Bereich der Morphologie”, in Lehfeldt, W. and Strauss, U. (eds.) *Glottometrika 4*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 95-102.

Geršić, S. and Altmann, G. (1980) “Laut – Silbe – Wort und das Menzerathsche Gesetz”, *Frankfurter phonetische Beiträge III = Forum Phoneticum*, 21, pp. 115-123.

Grégoire, A. (1899) *Variations de durée de la syllabe française suivant sa place dans les groupements phonétiques*, La Parole, Paris.

Grotjahn, R. (1992) “Evaluating the adequacy of regression models: Some potential pitfalls”, in Rieger, B. (ed.) *Glottometrika 13*, Universitätsverlag Dr. N. Brockmeyer, Bochum, pp. 121-172.

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. (1999) “Randbemerkungen zur Korrelation von Wort- und Silbenlänge im Kroatischen”, in Tošović, B. (ed.) *Die grammatischen Korrelationen*, Institut für Slawistik, Graz, pp. 67-77.

Grzybek, P. (2000) “Pogostnostna analiza based iz elektronskega korpusa slovenskih besedil”, *Slavistična revija*, 48(2), pp. 141-157, available at: https://srl.si/ojs/srl/article/view/2000-2-1-2, accessed: 30 April 2022.

Grzybek, P. (2010) “Text difficulty and the Arens-Altmann law”, in Grzybek, P., Kelih, E. and Mačutek, J. (eds.) *Text and Language: Structures · Functions · Interrelations Quantitative Perspectives*, Praesens Verlag, Wien, pp. 57-70.

Grzybek, P. (2013) “Close and distant relatives of the sentence: Some results from Russian”, in Obradović, I., Kelih, E. and Köhler, R. (eds.) *Methods and Applications of Quantitative Linguistics*, Akademska Misao, Beograd, pp. 59-68.

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.

Grzybek, P., Kelih, E. and Stadlober, E. (2008) “The relation between word length and sentence length: an intra-systemic perspective in the core data structure”, *Glottometrics*, 16, pp. 111-121, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g16zeit.pdf, accessed: 30 April 2022.

Grzybek, P., Stadlober, E. and Kelih, E. (2007) “The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective”, in Decker, R. and Lenz, H.-J. (eds.) *Advances in Data Analysis: Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, Berlin, Heidelberg, pp. 1-8, doi: 10.1007/978-3-540-70981-7_70.

Gustison, M. L. and Bergman, T. J. (2017) “Divergent acoustic properties of gelada and baboon vocalizations and their implications for the evolution of human speech”, *Journal of Language Evolution*, 2(1), pp. 20-36, doi: 10.1093/jole/lzx015.

Gustison, M. L. et al. (2016) “Gelada vocal sequences follow Menzerath’s linguistic law”, *Proceedings of the National Academy of Sciences*, 113(19), pp. E2750-E2758, doi: 10.1073/pnas.1522072113.

Hall, T. A. (2006) “Syllable: Phonology”, in Brown, K. (ed.) *Encyclopedia of Language & Linguistics*, 2nd edn, Elsevier, pp. 329-333.

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

Heesen, R. et al. (2019) “Linguistic laws in chimpanzee gestural communication”, *Proceedings of the Royal Society B: Biological Sciences*, 286(28620182900), doi: 10.1098/rspb.2018.2900.

Hernández-Fernández, A. et al. (2019) “Linguistic Laws in Speech: The Case of Catalan and Spanish”, *Entropy*, 21(1153), doi: 10.3390/e21121153.

Heups, G. (1983) “Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen”, in Köhler, R. and Boy, J. (eds.) *Glottometrika 5*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 113-133.

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. (1990a) “Menzerath-Altmann's law on the semantic level”, in Hřebíček, L. (ed.) *Glottometrika 11*, Universitätsverlag Dr. N. Brockmeyer, Bochum, pp. 47-56.

Hřebíček, L. (1990b) “The Constants of the Menzerath-Altmann Law”, in Hammerl, R. (ed.) *Glottometrika 12*, Universitätsverlag Dr. N. Brockmeyer, Bochum, pp. 61-71.

Hřebíček, L. (1994) “Fractals in language”, *Journal of Quantitative Linguistics*, 1(1), pp. 82-86, doi: 10.1080/09296179408590001.

Hřebíček, L. (1995) *Text Levels: Language Constructs, Constituents and the Menzerath-Altmann Law*, WVT Wissenschaftlicher Verlag Trier, Trier.

Hřebíček, L. (1997) *Lectures on Text Theory*, Oriental Institute, Prague.

Hřebíček, L. (2002a) “The elements of symmetry in text structures”, *Glottometrics*, 2, pp. 17-33, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g2zeit.pdf, accessed: 30 April 2022.

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

Huang, M. et al. (2020) “Male gibbon loud morning calls conform to Zipf’s law of brevity and Menzerath’s law: insights into the origin of human language”, *Animal Behaviour*, 160, pp. 145-155, doi: 10.1016/j.anbehav.2019.11.017.

Huang, S.-Z., Jin, J. and Shi, D. (2016) “Adjectives and adjective phrases”, in Huang, C.-R. and Shi, D. (eds.) *A Reference Grammar of Chinese*, Cambridge University Press, Cambridge, pp. 276-296, doi: 10.1017/CBO9781139028462.011.

Hug, M. (2004) “La loi de Menzerath appliquée à un ensemble de textes”, *Lexicometrica*, 5, pp. 1-10, available at: http://lexicometrica.univ-paris3.fr/article/numero5/lexicometrica-hug.pdf, accessed: 30 April 2022.

IDS UCS Basic (2022) CHaracter Information Service Environment (CHISE), available at: https://gitlab.chise.org/CHISE/ids/-/blob/master/IDS-UCS-Basic.txt, accessed: June 17, 2022.

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.

Just, M. A. and Carpenter, P. A. (1992) “A Capacity Theory of Comprehension: Individual Differences in Working Memory”, *Psychological Review*, 99(1), pp. 122-149.

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.

Kelih, E. (2012) “Systematic Interrelations Between Grapheme Frequencies and Word Length: Empirical Evidence from Slovene*”, *Journal of Quantitative Linguistics*, 19(3), pp. 205-231, doi: 10.1080/09296174.2012.685304.

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. (1989) “Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus”, in Altmann, G. and Schwibbe, M. H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Georg Olms Verlag, Hildesheim, pp. 108-112.

Köhler, R. (1993) “Synergetic Linguistics”, 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. 41-51.

Köhler, R. (1999) “Syntactic Structures: Properties and Interrelations”, *Journal of Quantitative Linguistics*, 6(1), pp. 46-57, doi: 10.1076/jqul.6.1.46.4137.

Köhler, R. (2002) “Power law models in linguistics: Hungarian”, *Glottometrics*, 5, pp. 51-61, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g5zeit.pdf, accessed: 30 April 2022.

Köhler, R. (2005) “Synergetic linguistics”, 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. 760-774.

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.

Kraviarová, M. and Zimmermann, J. (2010) “Menzerathov zákon v slovenskom vedeckom texte (Menzerath’s law in a Slovak scientific text)”, *Jazyk a kultúra (Language and Culture)*, 1, available at: https://www.ff.unipo.sk/jak/1_2010/kraviarova_zimmermann.pdf, accessed: 30 April 2022.

Krott, A. (1996) “Some remarks on the relation between word length and morpheme length”, *Journal of Quantitative Linguistics*, 3(1), pp. 29-37, doi: 10.1080/09296179608590061.

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

Kułacka, A. (2009a) “The necessity of the Menzerath-Altmann law”, *Anglistica Wratislaviensia*, 47, pp. 55-60, available at: https://wuwr.pl/awr/article/download/120/99/, accessed: 30 April 2022.

Kułacka, A. (2009b) “Warunki zachodzenia prawa Menzeratha-Altmanna”, *LingVaria*, 1(7), pp. 17-28.

Kułacka, A. (2010) “The Coefficients in the Formula for the Menzerath-Altmann Law”, *Journal of Quantitative Linguistics*, 17(4), pp. 257-268, doi: 10.1080/09296174.2010.512160.

Kułacka, A. and Mačutek, J. (2007) “A discrete formula for the Menzerath-Altmann law*”, *Journal of Quantitative Linguistics*, 14(1), pp. 23-32, doi: 10.1080/09296170600850585.

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.

Lehfeldt, W. (2007) “The Fall of the Jers in the Light of Menzerath’s Law”, in Grzybek, P. (ed.) *Contributions to the Science of Text and Language: Word Length Studies and Related Issues*, Springer, Dordrecht, pp. 211-213.

Lehfeldt, W. and Altmann, G. (2002) “Der altrussische Jerwandel”, *Glottometrics*, 2, pp. 33-44, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g2zeit.pdf, accessed: 30 April 2022.

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

Li, W. (2012) “Menzerath's law at the gene-exon level in the human genome”, *Complexity*, 17, pp. 49-53, doi: 10.1002/cplx.20398.

Li, Y.-H. A. (2016) “Verbs and verb phrases”, in Huang, C.-R. and Shi, D. (eds.) *A Reference Grammar of Chinese*, Cambridge University Press, Cambridge, pp. 81-115, doi: 10.1017/CBO9781139028462.005.

Luke, J. (2006) “Lun xiaoju zai Hanyu yufa zhong de diwei (On the status of clause in Chinese grammar)”, *Hanyu Xuebao (Journal of the Chinese language)*, 15(3), pp. 2-14.

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.

Marcus, M. et al. (1994) “The Penn Treebank: Annotating Predicate Argument Structure”, in *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, available at: https://aclanthology.org/H94-1020, accessed: 30 April 2022.

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.

Menzerath, P. and de Oleza, J. M. (1928) *Spanische Lautdauer. Eine experimentelle Untersuchung*, Walter de Gruyter, Berlin und Leipzig.

Meyer, P. (2002) "Laws and theories in quantitative linguistics", *Glottometrics*, 5, pp. 62-80, available at: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g5zeit.pdf, accessed: 30 April 2022.

Mikros, G. and Milička, J. (2014) "Distribution of the Menzerath's law on the syllable level in Greek texts", in Altmann, G. et al. (eds.) *Empirical Approaches to Text and Language Analysis*, RAM-Verlag, Lüdenscheid, pp. 1-10.

Milička, J. (2014) "Menzerath's Law: The Whole is Greater than the Sum of its Parts", *Journal of Quantitative Linguistics*, 21(2), pp. 85-99, doi: 10.1080/09296174.2014.882187.

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. and Schusterová, D. (2016) "Menzerath-Altmann Law – Analyses of Short Stories Written by Chinese Authors", in Benešová, M. (ed.) *Text segmentation for Menzerath-Altmann law testing*, Palacký University Olomouc, Olomouc, pp. 71-116.

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.

Mullaney, T. S. (2017) "Quote Unquote Language Reform: New-Style Punctuation and the Horizontalization of Chinese", *Modern Chinese Literature and Culture*, 29(2), pp. 206-250.

Nikolaou, C. (2014) "Menzerath–Altmann law in mammalian exons reflects the dynamics of gene structure evolution", *Computational Biology and Chemistry*, 53, pp. 134-143, doi: 10.1016/j.compbiolchem.2014.08.018.

Nivre, J. et al. (2016) "Universal Dependencies v1: A Multilingual Treebank Collection", in Calzolari, N. et al. (eds.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, European Language Resources Association (ELRA), Portorož, pp. 1659–1666, available at: https://aclanthology.org/L16-1262/, accessed: 30 April 2022.

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

Osborne, T. (2019b) "13. Dependency Grammar", in Kertész, A., Moravcsik, E. and Rákosi, C. (eds.) *Current Approaches to Syntax: A Comparative Handbook*, De Gruyter Mouton, Berlin, Boston, pp. 361-388, doi: 10.1515/9783110540253-013.

Osborne, T. and Gerdes, K. (2019) "The status of function words in dependency grammar: A critique of Universal Dependencies (UD)", *Glossa: a journal of general linguistics*, 4(1), pp. 1-28, doi: 10.5334/gjgl.537.

Parataxis: parataxis (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/parataxis.html, accessed: 30 April 2022.

Parataxis: parataxis (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/parataxis.html, accessed: 30 April 2022.

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.

Poiret, R. et al. (2021) “Universal Dependencies for Mandarin Chinese”, *Language Resources and Evaluation*, pp. 1-38, doi: 10.1007/s10579-021-09564-2.

Polikarpov, A. A. (2000) “Chronological Morphemic and Word-Formational Dictionary of Russian: Some System Regularities for Morphemic Structures and Units”, *Linguistische Arbeitsberichte*, 75, pp. 201-212.

Prün, C. (1994) “Validity of Menzerath-Altmann's law: Graphic representation of language, information processing systems and synergetic linguistics”, *Journal of Quantitative Linguistics*, 1(2), pp. 148-155, doi: 10.1080/09296179408590009.

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

Rettweiler, H. (1950) *Die Stichprobenentnahme bei sprachtypologischen Untersuchungen, als Problem nachgeprüft an der italienischen Sprache*, Ph.D. Dissertation.

Roberts, A. H. (1965) *A statistical linguistic analysis of American English*, Mouton, The Hague.

Rothe-Neves, R., Bernardo, B. M. and Espesser, R. (2017) “Shortening Tendency for Syllable Duration in Brazilian Portuguese Utterances”, *Journal of Quantitative Linguistics*, pp. 1-12, doi: 10.1080/09296174.2017.1360172.

Rothe, U. (1983) “Wortlänge und Bedeutungsmenge: Eine Untersuchung zum Menzerathschen Gesetz an drei romanischen Sprachen”, in Köhler, R. and Boy, J. (eds.) *Glottometrika 5*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 101-112.

Roukk, M. (2003a) *Стилистические характеристики речи учащихся 3-4 клас- сов (с применением математических методов) (The speech style characteristics of 3rd and 4th year school pupils (using mathematical methods))*, Ph.D. Dissertation.

Roukk, M. (2003b) “The Menzerath-Altmann Law in Russian Texts”, in *IV. Trierer Kolloquium zur Quantitativen Linguistik, 16.-18. Oktober 2003, Trier*.

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.

Rovenchak, A. (2015) “Quantitative Studies in the Corpus of Nko Periodicals”, in A. Tuzzi, M. Benešová and J. Mačutek (eds.) *Recent Contributions to Quantitative Linguistics*, De Gruyter Mouton, Berlin, München, Boston, pp. 125-138, doi: 10.1515/9783110420296-012.

Rujević, B. et al. (2021) “Quantitative analysis of syllable properties in Croatian, Serbian, Russian, and Ukrainian”, in Pawłowski, A. et al. (eds.) *Language and Text: Data, models,* *information and applications*, John Benjamins Publishing Company, Amsterdam, pp. 55-67, doi: 10.1075/cilt.356.04ruj.

Sambor, J. (1984) “Menzerath's law and the polysemy of words”, in Boy, J. and Köhler, R. (eds.) *Glottometrika 6*, Studienverlag Dr. N. Brockmeyer, Bochum, pp. 94-114.

Sanada, H. (2016) “The Menzerath-Altmann Law and Sentence Structure”, *Journal of Quantitative Linguistics*, 23(3), pp. 256-277, doi: 10.1080/09296174.2016.1169850.

Schusterová, D. et al. (2013) “An Application of the Menzerath-Altmann Law to Contemporary Spoken Chinese”, *Czech and Slovak Linguistic Review*, 1, pp. 55-73.

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.

Schwibbe, M. H. (1989) “Die Menzerathsche Regel als Modell psychischer Informationsverarbeitung”, in Altmann, G. and Schwibbe, M. H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Georg Olms Verlag, Hildesheim, pp. 84-91.

Ščigulinská, J. and Schusterová, D. (2014) *An Application of the Menzerath–Altmann Law to Contemporary Spoken Chinese*, Palacký University Olomouc, Olomouc.

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 organization 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)*.

Shi, D. (2016) “Nouns and nominal phrases”, in Huang, C.-R. and Shi, D. (eds.) *A Reference Grammar of Chinese*. Cambridge: Cambridge University Press, pp. 199-255, doi: 10.1017/CBO9781139028462.009.

Sievers, E. (1901) *Grundzüge der Phonetik: Zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen*, Breitkopf & Härte, Leipzig.

Simple Clauses (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/simple-syntax.html, accessed: 30 April 2022.

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

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

Strazny, P. (ed.) (2005) *Encyclopedia of linguistics*, Taylor & Francis, Oxon.

Subordination (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/complex-syntax.html#subordination, accessed: 30 April 2022.

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.

Syntax: General Principles (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/syntax.html, accessed: 30 April 2022.

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.

Tanaka-Ishii, K. (2021) "Menzerath's Law in the Syntax of Languages Compared with Random Sentences", *Entropy*, 23(661), pp. 1-12, doi: 10.3390/e23060661.

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.

The Primacy of Content Words (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/syntax.html#the-primacy-of-content-words, accessed: 30 April 2022.

Tokenization and Word Segmentation (2021), Universal Dependencies, available at: https://universaldependencies.org/u/overview/tokenization.html, accessed: 30 April 2022).

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.

Torre, I. G. et al. (2019) "On the physical origin of linguistic laws and lognormality in speech", *Royal Society Open Science*, 6(191023), doi: 10.1098/rsos.191023.

Tuldava, J. (1995) "Informational measures of causality*", *Journal of Quantitative Linguistics*, 2(1), pp. 11-14, doi: 10.1080/09296179508590028.

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

UD_Chinese-GSDSimp (2021), UniversalDependencies, available at: https://github.com/UniversalDependencies/UD_Chinese-GSDSimp, accessed: 30 April 2022.

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

UD_Chinese-HK (2021), UniversalDependencies, available at: https://github.com/UniversalDependencies/UD_Chinese-HK/tree/master, accessed: 30 April 2022.

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

UD_Chinese-PUD (2021), UniversalDependencies, available at: https://github.com/UniversalDependencies/UD_Chinese-PUD/tree/master, accessed: 30 April 2022.

Uhlířová, L. (1995) "O jednom modelu rozložení délky slov (On a Model of Word Length Distribution)", *Slovo a slovesnost (Word and Word Art)*, 56(1), pp. 8-14.

Unicode (2021) *Ideographic Description Characters: Range: 2FF0–2FFF*, available at: https://www.unicode.org/charts/PDF/U2FF0.pdf, accessed: 17 June 2022.

Universal Dependency Relations (2021), Universal Dependencies, available at: https://universaldependencies.org/u/dep/, 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.

Vulanović, R. and Köhler, R. (2005) "Syntactic units and structures", 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. 274-291.

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.

Ward, G. (2002) *Moby Hyphenation List*, available at: https://www.gutenberg.org/ebooks/3204, accessed: 30 April 2022.

Watson, S. K. et al. (2020) "An exploration of Menzerath's law in wild mountain gorilla vocal sequences", *Biology Letters*, 16(20200380), doi: 10.1098/rsbl.2020.0380.

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.

Wilde, M. H. and Schwibbe, J. (1989) "Organizationsformen von Erbinformation im Hinblick auf die Menzerathsche Regel", in Altmann, G. and Schwibbe, M. H., *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, Georg Olms Verlag, Hildesheim, pp. 92-107.

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.

Xcomp: open clausal complement (2021a), Universal Dependencies, available at: https://universaldependencies.org/u/dep/xcomp.html, accessed: 30 April 2022.

Xcomp: open clausal complement (2021b), Universal Dependencies, available at: https://universaldependencies.org/zh/dep/xcomp.html, accessed: 30 April 2022.

Xia, F. (2000) "The Segmentation Guidelines for the Penn Chinese Treebank (3.0)". University of Pennsylvania Institute for Research in Cognitive Science Technical Report No. IRCS-00-06, available at: https://repository.upenn.edu/ircs_reports/37/, accessed: 30 April 2022.

Xue, N. et al. (2013) *Chinese Treebank 8.0 LDC2013T21*, Linguistic Data Consortium, Philadelphia, doi: 10.35111/wygn-4f57.

Xu, L. and He, L. (2018) "Is the Menzerath-Altmann Law Specific to Certain Languages in Certain Registers?", *Journal of Quantitative Linguistics*, pp. 1-17, doi: 10.1080/09296174.2018.1532158.

Zaliznjak, A. A. (1985) *Ot praslavjanskoj akcentuacii k russkoj (From Proto-Slavic stress to Russian)*, Nauka, Moskva.

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. (2021a) *Universal Dependencies 2.8.1*, 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/11234/1-3687, accessed: 30 April 2022.

Zeman, D. et al. (2021b) *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 Tables

List of Figures