The Menzerath-Altmann law in syntactic structure revisited: Combining linearity of language with dependency syntax

Ján Mačutek, Radek Čech, Marine Courtin
Mathematical Institute, Slovak Academy of Sciences, Slovakia & Department of Mathematics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Slovakia
Department of Czech Language, Faculty of Arts, University of Ostrava, Czech Republic
LPP (CNRS) – Sorbonne Nouvelle, France
jmacutek@yahoo.com, cechradek@gmail.com, marine.courtin@sorbonne-nouvelle.fr

Abstract

According to the Menzerath-Altmann law, there is inverse proportionality between sizes of language units and their constituents (i.e., longer language units are composed of shorter constituents, and vice versa). The validity of the law was confirmed many times for the relation between lengths of a word and its syllables. However, the relation between lengths of sentences (measured in clauses) and clauses (measured in words) is problematic. In this paper, a new language unit – linear dependency segment – is introduced with the motivation to avoid some problems connected to the Menzerath-Altmann law on the syntactic level. The new unit is intermediate between clause and word and its definition takes into account both the linearity of language and dependency syntactic structure. It is shown that the relation between sentence length in clauses and clause length measured in linear dependency segments abides by the Menzerath-Altmann law in two Czech dependency treebanks.

1 Introduction

The Menzerath-Altmann law (MAL henceforward) predicts relations between sizes of language units which are neighbours in the language unit hierarchy. According to the law, longer units which are higher in the hierarchy (constructs) consist of shorter lower units (constituents). The formulation of the MAL developed from a verbal one (the longer the word, the shorter on average its syllables; see Menzerath, 1954) to mathematical formula

$$y(x) = ax^b e^{-cx} \quad (1)$$

derived by Altmann (1980). In formula (1), $y(x)$ is the mean size of constituents in the construct of size x; a, b and c are parameters. Very often a simpler formula,

$$y(x) = ax^b, \quad (2)$$

is used, which is a special case of function (1) for $c = 0$.

The MAL was first observed as the relation between word length in syllables and either syllable length in phonemes[1] (Menzerath, 1954), or syllable duration in time (Menzerath and de Oleza, 1928;

[1] Sometimes, word length is measured in graphemes instead of phonemes. This approach is applied mainly in languages which have a close phoneme-grapheme correspondence.

Geršić and Altmann, 1980). The validity of the MAL at this lowest level was scrutinized in many languages (see e.g. Cramer, 2005, and references therein; Kelih, 2010, 2012; Mikros and Milička, 2014; Mačutek et al., 2019).

However, two fundamental problems emerge when one goes higher in the hierarchy of language units. First, it was assumed that the upper neighbours of word are clause and sentence. Although several papers in 1980s (Köhler, 1982; Heups, 1983; Schwibbe, 1984; Teupenhayn and Altmann, 1984) claim that the relation between sentence length in clauses and clause length in words abides by the MAL, more recent results are far from clear. Thus, Kułacka (2010), Chen and Liu (2019), and Xu and He (2020) confirm the older results, while data analysed by Kułacka and Mačutek (2007), Benešová and Čech (2015), and Hou et al. (2017) display a Menzerathian tendency, but they cannot be fitted by function (1) sufficiently well.[2] On the other hand, data presented by Buk and Rovenchak (2008) and by Andres and Benešová (2012) do not confirm to the MAL.[3] Curiously enough, Andres and Benešová (2012) and Hou et al. (2019) are, to our best knowledge, the only two papers which focus also on the relation between lengths of clause (in words) and word (in syllables).[4] This relation, again, cannot be modelled by the MAL. To put it mildly, the empirical evidence of the MAL, especially in form of function (2), is doubtful as soon as we move from word to clause and sentence.

Mačutek et al. (2017) tried to measure clause length in syntactic phrases which are directly dependent on the predicate of the clause (with phrase length being measured in words). The MAL in form (2) achieved a very good fit. The phrase thus became a candidate for an intermediate language unit between word and clause. It must be noted that only main clauses were analysed, and only one Czech treebank was used.

Second, although the linguistic interpretation of the parameters of model (1) is still not known, it was suggested that the MAL has something to do with short term memory (Köhler, 1989; Grzybek, 2013; see also Yngve, 1960, 1996).[5] According to the well-known paper by Miller (1956), the capacity of short-term memory is approximately seven. With the exception of polysynthetic languages, words only seldom contain more than seven syllables (or morphemes[6]), and the same is true for sentence length in clauses. However, clauses longer than seven words are not so rare – the mean clause length in the papers cited above is often somewhere near 10, see e.g. Köhler (1982), Heups (1983), and Teupenhayn and Altmann (1984).

The phrase used by Mačutek et al. (2017) faces the same problem, e.g. 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). In addition, consider a sentence consisting only of a single predicate (e.g. Czech sentence *Prší "It rains"*). Such a sentence contains only one clause of length zero (because there is nothing directly dependent on the predicate of the clause), and phrase length cannot be determined at all, as there is no phrase in the sense of the phrase definition from Mačutek et al. (2017). If the definition is modified so that phrase includes also the predicate, the question arises how to determine phrase length in clauses consisting of at least two phrases (such as e.g. in Czech sentence *Petr miluje Marii "Peter loves Mary"*). If the predicate is a part of the phrases, it appears more than once in all calculations. Regardless of these methodological difficulties, phrase has also a drawback of neglecting the linearity of language.

[2] See Mačutek and Wimmer (2013) for an overview of goodness-of-fit criteria usually used in quantitative linguistics.

[3] Admittedly, these papers do not follow the same methodology. In most of them, either finite verbs or punctuation marks (comma and semicolon) to determine sentence length in clauses.

[4] Hou et al. (2019) measure word length in characters, but in written Chinese there is almost one-to-one correspondence between characters and syllables.

[5] Torre et al. (2019) present an attempt to explain the origin of the MAL in spoken language at the level of words and syllables as a consequence of human physiology (in particular the necessity to breathe). These two tentative explanations of the MAL do not exclude each other; rather, both factors (pauses caused by breathing and a limited capacity of short-term memory) are likely to contribute to the shortening of constituents in longer constructs.

[6] See Pelegrinová et al. (2021) and references therein for the MAL as the relation between word length in morphemes and morpheme length in phonemes.

To avoid the abovementioned problems, we suggest another approach, namely, a new language unit between word and clause is introduced. Its definition combines both linear and hierarchical dependency structure of sentence. We focus on the question whether this new unit behaves according to the MAL.

The paper is structured as follows. Section 2 introduces the linear dependency segment, a new unit positioned between clause and word. In Section 3, language material used for the analysis is described. Results achieved are presented in Section 4. The paper is concluded by a short discussion which contains also some ideas for future research in this area.

2 Linear dependency segment

We define the linear dependency segment (LDS henceforward) as the longest possible sequence of words (belonging to the same clause[7]) 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). Figure 1 presents the dependency tree of sentence *"This black book on the table costs twenty euros, which is too much for me"*.

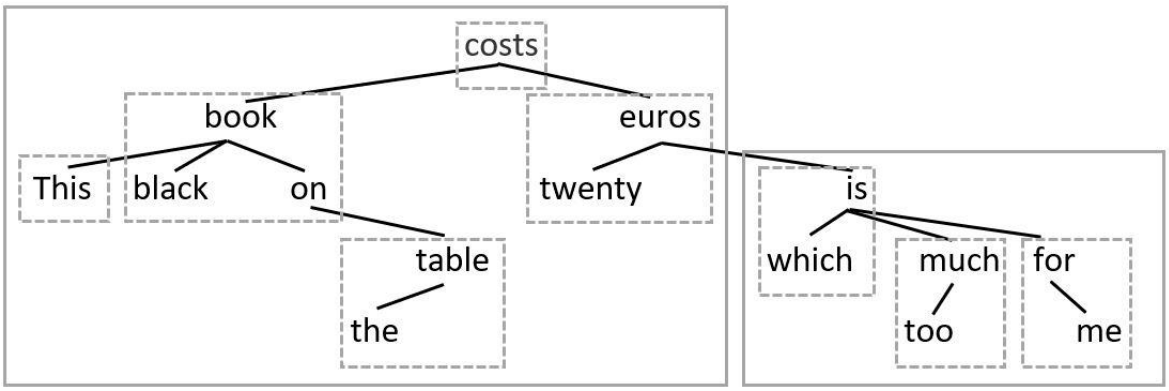

Figure 1. Dependency tree of sentence *"This black book on the table costs twenty euros, which is too much for me"*

Consider the first clause in the sentence. Its first word, "*This*", is syntactically linked with "*book*", but these two words are not linear neighbours. Therefore, the first LDS is [This]. Next, the second word, "*black*", is syntactically linked with "*book*", which is also its linear neighbour, and the third and the fourth words, "*book*" and "*on*", are again both linear and syntactic neighbours. Here the segment is ended, because the next word, "*the*", is not syntactically linked with "*on*". Examining the whole clause we obtain LDSs [This][black book on][the table][costs][twenty euros]. Similarly, the second clause in this sentence has LDSs [which is][too much][for me]. We remind that we define the LDSs as units of which clauses are composed, i.e. a LDS is always ended at the end of a clause.

The definition is good in the sense that every clause can be unambiguously divided into LDSs, and that the intersection of two different LDSs is the empty set (i.e. every word in a clause belongs to one and only one LDS).

From the MAL point of view, clause is a construct and LDS its constituent (which, in turn, is a construct itself, with words being its constituents). We expect that longer sentences (measured in the number of clauses) contain shorter clauses (measured in the number of LDSs), and vice versa. This expectation is based on the fact that dependency links which do not respect the linearity of a sentence are more difficult to process.[8] The same is true for a sentence with many clauses. The MAL does not allow sentences to become too complex, as it "forces" clauses in long sentences (i.e. in ones which

[7] We use the definition of clause from Prague Dependency Treebank 3.0 (https://ufal.mff.cuni.cz/pdt3.0/documentation#__RefHeading__42_1200879062), according to which "[a] clause typically corresponds to a single proposition expressed by a finite verb and all its arguments and modifiers (unless they constitute clauses of their own)".

[8] The idea that dependency distance in language is shorter than a random baseline can be traced back to Liu (2008).

contain many clauses) to become shorter (i.e. to be composed of fewer LDSs). Fewer LDSs mean that there are fewer dependency distances (as defined by Liu, 2008, p. 164) longer than one (as all dependency distances within one LDS are minimal, i.e. equal to one).

Provided that the MAL is valid as a model for the relation between lengths of sentences and clauses, a sentence can be composed either of more clauses which are shorter in terms in LDSs (which means that they are syntactically simpler[9]), or of fewer clauses which are "allowed" to contain more LDSs (and consequently to be syntactically more complex)

3 Language material

For the analysis, we used two Czech treebanks, the Czech-PDT UD[10] and the FicTree (Jelínek, 2017). The treebanks were converted to the Surface Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018). The use of the Universal Dependency annotation scheme (de Marneffe et al., 2021) was also considered. However, we prefer the SUD approach because it is based on surface-syntactic distributional criteria that fit the nature of our analysis better than the Universal Dependency approach which is based on "a mixture of semantic and syntactic motivations" (Osborne and Gerdes, 2019).

The Czech-PDT UD consists of 87,913 Czech sentences from non-abbreviated newspaper, business and popular scientific journal articles published from 1991 to 1995. The FicTree consists of 12,760 sentences from Czech literary works published between 1991 and 2007. The treebanks were also merged and treated as one whole in which different genres are represented. Sentences without a predicate (especially titles of newspaper articles) were removed. We thus analysed altogether 86,266 sentences.

4 Results

As we study the relation between sentence length and the mean clause length, the number of clauses from which the mean is calculated cannot be too low if the result should be robust. We decided to take into account sentence lengths with frequencies which make at least 0.1% of our language material. We thus disregarded sentences containing more than eight clauses (together 76 sentences, i.e. 0.09%). Very complicated structures, such as several clauses placed in brackets, clauses separated by a colon, or citations, are typical for these long sentences. The possibility to check thoroughly sentences which do not conform to the MAL was also the reason why we focus only on Czech treebanks in this paper – one of the coauthors is a native Czech speaker. It is obvious that our choice substantially limits the scope of this paper, but given that it is the first attempt to study the LDS as a language unit, we prefer this more careful approach.

The relation between sentence length in clauses and the mean clause length measured in LDSs is presented in Table 1.

[9] If we consider the extreme case, a clause consisting of only one LDS either contains only one word, or it reaches the minimum of dependency distance (in such a clause all dependency distances are equal to one).

[10] https://universaldependencies.org/treebanks/cs_pdt/index.html

	merged			PDT			FicTree		
SL	f	rf	MCL	f	rf	MCL	f	rf	MCL
1	36559	0.424	5.02	32002	0.428	5.30	4557	0.396	3.03
2	27735	0.321	3.93	24121	0.323	4.10	3614	0.314	2.82
3	13463	0.156	3.44	11605	0.155	3.54	1858	0.162	2.79
4	5416	0.063	3.17	4537	0.061	3.25	879	0.076	2.77
5	1962	0.023	3.00	1616	0.022	3.07	346	0.030	2.69
6	727	0.008	2.94	580	0.008	3.02	147	0.013	2.64
7	236	0.003	2.84	188	0.003	2.85	48	0.004	2.82
8	92	0.001	2.79	69	0.001	2.93	23	0.002	2.36

Table 1. The MAL in Czech dependency treebanks (SL - sentence length in clauses, f, rf - frequencies and relative frequencies[11] of sentence lengths, MCL – the mean clause length in LDSs).

The MAL in form (2) fits the data from the merged treebanks very well[12], with $R^2 = 0.9836$ ($a = 4.918$, $b = -0.296$).[13] The data and the graph of the function can be seen in Figure 2.

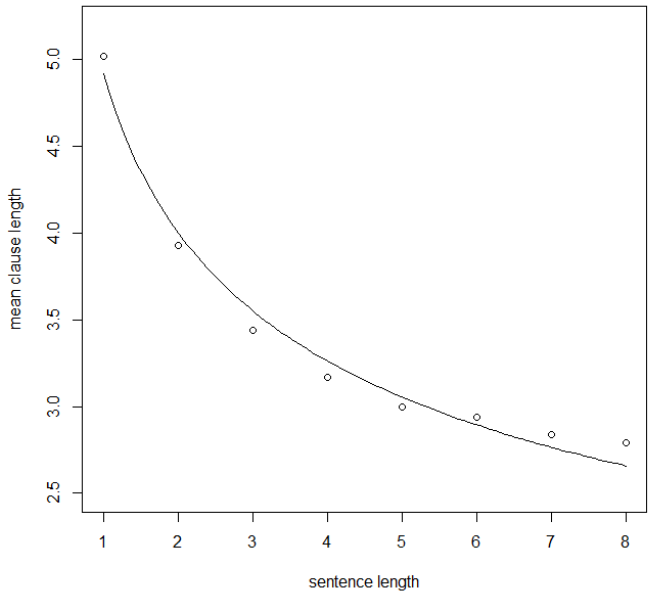

Figure 2. The MAL modelled by function $y(x) = ax^b$ as the relation between sentence length and the mean clause length

The value of parameter a is very close to the mean clause length (measured in the number of LDSs) in sentences consisting of only one clause. If we use this value, i.e. if we set $a = 5.02$ in formula (2),

[11] The relative frequencies do not sum to one, because sentences containing more than eight clauses were disregarded.

[12] The most common rule of thumb in quantitative linguistics is to consider the goodness-of-fit of a model satisfactory if the value of the determination coefficient R^2 is higher than 0.9, see Mačutek and Wimmer (2013).

[13] The fit remains satisfactory also if other options how to deal with low frequency construct length are applied. If all construct lengths with frequency at least 10 are used in the computations (see Mačutek and Rovenchak, 2011), we have $R^2 = 0.9353$, and if we pool low-frequency construct lengths (i.e. sentence which contain more than eight clauses in our case) and compute the weighted mean of clause lengths (see e.g. Pelegrinová et al., 2021), we obtain $R^2 = 0.9649$.

we obtain $b = -0.309$ and $R^2 = 0.9803$, which is still a very good fit. We thus have a very clear interpretation of the parameter a.[14] As for parameter b, its linguistic interpretation remains an open question.

In both PDT and FIC treebanks, the decreasing tendency of the mean clause length can be observed. While the fit of function (2) remains very good ($R^2 = 0.9739$) for PDT, it is much worse ($R^2 = 0.6148$) for the data from the FicTree treebank. However, this is caused by an irregular behaviour of the mean clause length of the two highest values of sentence length, which occur with relatively low frequencies (moreover, the FicTree treebank is much smaller than PDT), and an overall decreasing tendency can be seen also in results from this treebank.

The two treebanks differ also in the mean values of the shortest sentences (i.e. the ones containing only one clause). Most likely, it is a consequence of different sentence length distributions in the treebanks (the mean values are 1.97 for PDT and 2.11 for FicTree; see also relative frequencies of sentence lengths in Table 1). Longer sentences in FicTree are composed of shorter LDSs. We remind that the treebanks consist of journalistic texts (PDT) and fiction (FicTree), and that sentence length depends on genre (see e.g. Kelih et al., 2006; Xu and He, 2020).

5 Conclusion

The achieved results indicate that, at least tentatively, the LDS can be considered a meaningful linguistic unit which allows to model the MAL also on the level of syntax. The LDS avoids the problems frequently encountered when one measures clause length in the number of words the clause contains. From the theoretical point of view, it is important that clause length measured in LDSs correspond with the capacity of short-term memory[15], which is one of theoretical explanations of the MAL. Furthermore, we emphasize that the definition of the LDS takes into account both the linearity of language and the dependency syntactic structure.

Naturally, this paper is only a pilot study, very limited in its scope, and data from many more typologically diverse languages must be analysed before the LDS can establish itself firmly among more traditional language units. Specifically with respect to the MAL, also relations between lengths of clauses (in LDSs) and LDSs (in words) and between lengths of LDSs (in words) and words (in syllables or morphemes) must be investigated. In addition, if the LSD turns out to be a suitable linguistic unit, also its frequencies and its length are supposed to follow distribution laws which are commonly used to model these language properties (i.e. a Zipf-like distribution for LDS frequencies, and a Poisson-like distribution for LSD length, see e.g. Popescu et al., 2009, and Grzybek, 2006, respectively).

Parameter values of the MAL in form of function (2) can probably be used in automatic text classification procedures, as they depend on sentence length, which, in turn, depends on genre.

A possible correspondence between LDSs and dependency distance minimization deserves a closer inspection. While there is a strong evidence that words which are syntactically linked are close to each other also with respect to the linear order of the sentence (see e.g. Liu, 2008; Ferrer-i-Cancho and Liu, 2014; Futrell et al., 2015), short sentences are quite likely not to follow this trend (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). Although sentence length in these studies is expressed in the number of words (as opposed to clauses from our approach) they contain, we can suppose that short sentences mostly contain one or two clauses. The MAL predicts that clauses in short sentences are composed of relatively many LDSs, which means that there must be relatively many dependency distances with values more than one. The findings from Ferrer-i-Cancho and Gómez-Rodríguez (2021) and from this paper thus support each other.

[14] The interpretation of parameter a of the MAL in form (2) as the mean length of constituents of the shortest constructs is not specific to language units analysed in this paper – e.g. Kelih (2010) uses the same approach when investigating the relations between lengths of words and syllables.

[15] Miller (1956) claims that the capacity is roughly seven (although there are also other opinions). Clause length determined in the number of the LDSs only rarely exceeds this value, while clause length in words can be, naturally, (much) higher. Similarly, phrases used by Mačutek et al. (2017) contain more words than LDSs; in addition, the methodology from that paper allows to analyse only main clauses.

Acknowledgements

J Mačutek was supported by research grant VEGA 2/0096/21.

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. In Rüdiger Grotjahn, editor, *Glottometrika 2*, pages 1–10. Brockmeyer, Bochum.

Jan Andres and Martina Benešová. 2012. Fractal Analysis of Poe's Raven, II. *Journal of Quantitative Linguistics*, 19(4):301–324.

Martina Benešová and Radek Čech. 2015. Menzerath-Altmann law versus random model. In George K. Mikros and Ján Mačutek, editors, *Sequences in Language and Text*, pages 57–69. de Gruyter, Berlin / New York.

Solomija Buk and Andrij Rovenchak. 2008. Menzerath–Altmann law for syntactic structures in Ukrainian. *Glottotheory*, 1(1):10–17.

Heng Chen and Haitao Liu. 2019. A quantitative probe into the hierarchical structure of written Chinese. In Xinying Chen and Ramon Ferrer-i-Cancho, editors, *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 83–88. ACL, Stroudsburg (PA).

Irene M. Cramer. 2005. Das Menzerathsche Gesetz. In Reinhard Köhler, Gabriel Altmann, and Rajmund G. Piotrowski, editors, *Quantitative Linguistics. An International Handbook*, pages 659–688. de Gruyter, Berlin / New York.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. (2021).Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Ramon Ferrer-i-Cancho and Carlos Gómez-Rodríguez, 2021. Anti dependency distance minimization in short sequences. A graph theoretic approach. *Journal of Quantitative Linguistics*, 28(1):50–76.

Ramon Ferrer-i-Cancho and Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 5(2):143–355.

Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences of the United States of America*, 112(33):10336–10341.

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In Marie-Catherine de Marneffe, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies Workshop (UDW 2018)*, pages 66–74. ACL, Stroudsburg (PA).

Slavko Geršić and Gabriel Altmann. 1980. Laut – Silbe – Wort und das Menzerathsche Gesetz. In Hans-Walter Wodarz, editor, *Frankfurter phonetische Beiträge 3*, pages 115-123. Buske, Hamburg.

Peter Grzybek. 2006. History and methodology of word length studies. In: Peter Grzybek, editor, *Contributions to the Science of Language and Text. Word Length Studies and Related Issues*, pages 15–90. Dordrecht, Springer.

Peter Grzybek. 2013. Close and distant relatives of the sentence: Some results from Russian. In: Ivan Obradović, Emmerich Kelih, and Reinhard Köhler, editors, *Methods and Applications of Quantitative Linguistics*, pages 59-68. Beograd: Akademska Misao.

Gabriela Heups. 1983. Untersuchungen zum Verhältnis von Satzlänge zu Clauselänge am Beispiel deutscher Texte verschiedener Textklassen. In Reinhard Köhler and Joachim Boy, editors, *Glottometrika 5*, pages 113-133. Brockmeyer, Bochum.

Renkui Hou, Chu-Ren Huang, Hue San Do, and Hongchao Liu. 2017. A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 24(4):350–366.

Renkui Hou, Chu-Ren Huang, Mi Zhou, and Menghan Jiang. 2019. Distance between Chinese registers based on the Menzerath-Altmann law and regression analysis. *Glottometrics*, 45:24–57.

Tomáš Jelínek 2017. FicTree: A manually annotated treebank of Czech fiction. In Jaroslava Hlaváčová, editor, *Proceedings of the 17th Conference on Information Technologies – Applications and Theory (ITAP 2017)*, pages 181–185. http://ceur-ws.org/Vol-1885/181.pdf

Emmerich Kelih. 2010. Parameter interpretation of Menzerath law: Evidence from Serbian. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek, editors, *Text and language. Structures, functions, interrelations, quantitative perspectives*, pages 71–79. Praesens, Wien.

Emmerich Kelih. 2012. Systematic interrelations between grapheme frequencies and words length: Empirical evidence from Slovene. *Journal of Quantitative Linguistics*,19(3):205–231.

Emmerich Kelih, Peter Grzybek, Gordana Antić, and Ernst Stadlober. 2006. Quantitative text typology. The impact of sentence length. In Myra Spiliopoulou, Rudolf Kruse, Christian Borgelt, Andreas Nürnberger, and Wolfgang Gaul, editors, *From Data and Information Analysis to Knowledge Engineering*, pages382–389. Springer, Heidelberg / Berlin.

Reinhard Köhler. 1982. Das Menzeratsche Gesetz auf Satzebene. In Werner Lehfeldt and Udo Strauss, editors, *Glottometrika 4*, pages 103–113. Brockmeyer, Bochum.

Reinhard Köhler. 1989. Das Menzerathsche Gesetz als Resultat des Sprachverarbeitungsmechanismus. In Gabriel Altmann and Michael H. Schwibbe, editors, *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*, pages 108–112. Olms, Hildesheim / Zürich / New York.

Reinhard Köhler. 2012. *Quantitative Syntax Analysis*. de Gruyter, Berlin / Boston.

Agnieszka Kułacka. 2010. The coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 17(4):23–32.

Agnieszka Kułacka and Ján Mačutek. 2007. A discrete formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics*, 14(1):257–268.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

Ján Mačutek, Jan Chromý, and Michaela Koščová. 2019. Menzerath-Altmann law and prothetic /v/ in spoken Czech. *Journal of Quantitative Linguistics*, 26(1):66–80.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni and Joakim Nivre, editors, *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100-107. Linköping University Electronic Press: Linköping.

Ján Mačutek and Andrij Rovenchak. 2011. Canonical word forms: Menzerath-Altmann law, phonemic length and syllabic length. In Emmerich Kelih, Victor Levickij, and Yuliya Matskulyak, editors, *Issues in Quantitative Linguistics 2*, pages 136-147. RAM-Verlag, Lüdenscheid.

Ján Mačutek and Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20(3):227–240.

Paul Menzerath. 1954. *Die Architektonik des deutschen Wortschatzes*. Dümmler, Bonn.

Paul Menzerath and José M. de Oleza. 1928. *Spanische Lautdauer. Eine experimentelle Untersuchung.* de Gruyter, Berlin / Leipzig.

George Mikros and Jiří Milička. 2014. Distribution of the Menzerath's law on the syllable level in Greek texts. In Gabriel Altmann, Radek Čech, Ján Mačutek, and Ludmila Uhlířová, editors, *Empirical Approaches to Text and Language Analysis*, pages 181–189. RAM-Verlag, Lüdenscheid,

George A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

Timothy Osborne and Kim Gerdes. 2019. The status of function words independency grammar: A critique of Universal Dependencies (UD). *Glossa: a journal of general linguistics*, 4(1):17.

Kateřina Pelegrinová, Ján Mačutek, and Radek Čech. 2021. The Menzerath-Altmann law as the relation between lengths of words and morphemes in Czech. *Jazykovedný časopis*, 72 (to appear).

Ioan-Iovitz Popescu, Gabriel Altmann, Petr Grzybek, Bijapur D. Jayaram, Reinhard Köhler, Viktor Krupa, Ján Mačutek, Regina Pustet, Ludmila Uhlířová, and Matummal N. Vidya. 2009. *Word Frequency Studies*. de Gruyter, Berlin / New York.

Michael H. Schwibbe. 1984. Text- und wortstatistische Untersuchungen zur Validität der Menzerath'schen Regel. In Joachim Boy and Reinhard Köhler, editors, *Glottometrika 6*, pages 127–138. Brockmeyer, Bochum.

Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath's law. In Joachim Boy and Reinhard Köhler, editors, *Glottometrika 6*, pages 152–176. Brockmeyer, Bochum.

Iván G. Torre, Bartolo Luque, Lucas Lacasa, Christopher T. Kello, and Antoni Hernández-Fernández. 2019. On the physical origin of linguistic laws and lognormality in speech. *Royal Society Open Science*, 6:191023.

Lirong Xu and Lianzhen He. 2020. Is the Menzerath-Altmann law specific to certain languages in certain registers? *Journal of Quantitative Linguistics*, 27(3):187–203.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104(5):444 –446.

Victor H. Yngve. 1996. *From Grammar to Science. New Foundations for General Linguistics*. Benjamins, Amsterdam / Philadelphia.