Syntactic units and their length distributions: A case study in Czech

Michaela Nogolová[1], Michaela Koščová[2], Ján Mačutek[2], Radek Čech[3],

[1]Department of Czech Language, University of Ostrava
[2]Mathematical Institute, Slovak Academy of Sciences
[3]Department of Czech Language, Masaryk University
nogolovam@gmail.com, koscmichaela@gmail.com, jmacutek@yahoo.com, cechradek@gmail.com

Abstract

This study investigates the length distributions of syntactic units in Czech across multiple hierarchical levels: sentences, independent clauses, clauses, phrases, subphrases, chunks, and words. Using a diverse dataset -– including Universal Dependency treebanks, presidential speeches, the Czech Bible, and random sample from corpora of modern Czech – the analysis examines whether lengths of these syntactic units follow consistent distributional patterns. Length is defined as the number of immediate subunits, and the distributions were modeled using the right-truncated hyper-Poisson distribution. The results demonstrate that this model fits well distributions of length of all abovementioned syntactic units, pointing to a common principle underlying the organization of syntactic structure in Czech.

1 Introduction

The relationship between linguistic units has long been an important topic in quantitative linguistics, particularly through the study of the Menzerath-Altmann law (MAL henceforward, Menzerath, 1954; Altmann, 1980). The law expresses the relationship between the length of a construct and the length of its constituents (parts of the construct); specifically, the longer the construct, the shorter the constituent on average. While the MAL has been extensively validated using words and syllables, its applicability at the syntactic level remains a subject of ongoing investigation (Andres and Benešová, 2012, Sanada, 2016, Berdicevskis, 2021, Mačutek et al., 2017, Mačutek et al., 2021).

Recent study (Nogolová et al., 2025) has explored the MAL across several units in the language unit hierarchy (sentences - independent clauses – clauses – phrases – subphrases – chunks – words – syllables). These investigations suggest that the MAL holds across all these levels, indicating a consistent pattern also with respect to syntactic constructions. It is the first paper that considers several neighbouring language units simultaneously; all previous papers on this topic limited themselves to partial results, mostly only to one triad of units (such as, e.g., Mačutek et al., 2017 focus solely on clause length in phrases and phrase length in words).

Building upon this foundation, the present paper aims to analyze distributions of lengths of syntactic units mentioned in the previous paragraph. Length of a unit is measured in the number of its direct lower neighbours in the language unit hierarchy. Specifically, we will examine length of sentences (measured in the number of independent clauses the sentence contains), of independent clauses (in clauses), of clauses (in phrases), of phrases (ecubphrases), of subphrases (in chunks), of chunks (in words), and of words (in syllables). The goal is to determine whether length distributions of abovementioned syntactic units display similar patterns that might provide insight into the structure and organization of language units.

2 Language material

The language material for this study is sourced from multiple treebanks and corpora, each offering a distinct representation of the Czech language across various genres and time periods.

A significant portion is drawn from the Universal Dependencies 2.13 (Zeman et al., 2023). Specifically, we utilize six Czech dependency treebanks:

1. UD_Czech-CAC is based on the Czech Academic Corpus 2.0 (Vidová Hladká et al., 2008). This treebank encompasses articles from diverse sources, including journalism, administration, and scientific fields.

2. UD_Czech-CLTT originates from the Czech Legal Text Treebank 2.0 (Kríž and Hladká, 2017). It comprises two legal documents on accounting.

Proceedings of the Third Workshop on Quantitative Syntax (QUASY, SyntaxFest 2025), pages 115–123
August 29, 2025 ©2025 Association for Computational Linguistics

3. FicTree (Jelínek, 2017) includes six books of fiction, one book of fiction for children, and one memoir.

4. Czech-PDT UD (Bejček et al., 2013) contains journalistic texts.

5. UD_Czech-Poetry contains samples from 19th-century Czech poetry from the Corpus of Czech Verse (Plecháč and Kolár, 2015).

6. The Czech part of Parallel Universal Dependencies (PUD) consists of 1000 random sentences translated into Czech from English and other languages.[1]

In addition to the UD treebanks mentioned above, we analyze 89 annual speeches delivered by twelve Czechoslovak and Czech presidents. These speeches are the object of research in Kubát et al. (2021).

Our study also incorporates the Czech Ecumenical Translation of the Bible (CET), a contemporary Czech translation undertaken between 1961 and 1979. This translation renders biblical texts into modern Czech while preserving traditional diction and style. We use the 2001 revision of the CET translation.

Finally, we utilize sentences from the SYN2020 corpus (Křen et al., 2020), a comprehensive and balanced collection of contemporary written Czech developed by the Czech National Corpus. Predominantly encompassing texts from 2015 to 2019, the corpus contains 100 million words and is structured into three equally sized segments: fiction, non-fiction, and newspapers/magazines. For this study, a random sample of 50,000 sentences from each segment was used.

The individual parts were merged and treated as a whole, thus encompassing various genres. The language material used is quite heterogeneous. Menzerath (1954) formulated the relation between word length (in syllables) and the mean syllable length (in phonemes) as valid for the vocabulary. Similarly, we suppose that the Menzerath-Altmann law in general is valid for the inventory of units rather than for particular texts or text genres. Of course no dictionary or corpus contains the whole vocabulary, and it is even less realistic to speak about the complete inventory of sentences, clauses, etc., but some measure of heterogeneity reduces the risk of genre-specific syntactic units.

3 Methodology and operationalization

In this section, we present language units we analyze, following the approach from Nogolová et al. (2025).

Those parts of our corpus that do not come from UD treebanks (i.e., the presidential speeches, the Bible translation, and the sentences from the SYN2020 corpus; see Section 2) were processed using UDPipe 2.0 (Straka, 2018), a trainable pipeline that, among other functions, performs dependency parsing. Subsequently, all the annotated texts were converted to the Surface Syntactic Universal Dependencies (SUD) annotation scheme (Gerdes et al., 2018) using Grew software (Guillaume, 2021). The reason is that the SUD annotation reflects more closely dependency syntax based on purely syntactic (rather than semantic or functional) criteria. SUD provides a representation closer to surface-syntactic frameworks such as Meaning-Text Theory (Mel'čuk, 1988), Word Grammar (Hudson, 2010), and the Prague Dependency Treebank (Hajič et al., 2017).

We took sentences as they are determined by the annotation tool. Only sentences satisfying the following three criteria were included in the analysis: (i) they contain a predicate (a finite verb or an auxiliary) as the sentence root; (ii) they do not contain abbreviations, digits, foreign words, words with unknown syntactic functions, special characters, and words assigned the *flat*[2] or *orphan*[3] syntactic function; (iii) they do not contain a chain of more than two coordinated words, with exception of predicates of independent clauses (discussed below).

Sentence length is expressed as the number of independent clauses the sentence contains. The first independent clause comprises the root of the sentence and all words that are directly or indirectly syntactically linked with it if the links are not coordinations with another predicate. If there is another predicate coordinated with the sentence root, it becomes the root of another independent clause within the sentence.

The immediate constituent of an independent clause is a clause, which consists of the predicate

[1] These sentences form a part of the Parallel Universal Dependencies (PUD) treebanks created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies (http://universaldependencies.org/conll17/).

[2] Personal names, which exhibit a distinct syntactic behavior, are tagged by this relation.

[3] This often indicates ellipsis, which presents analytical challenges.

and all its direct or indirect dependents, excluding other predicates. To illustrate these units[4], Figure 1 presents the dependency tree of the sentence (1) divided into individual independent clauses (on the left) and clauses (on the right).

(1) *Ani u uvědomované motivace nemůžeme mít přehled o všech motivech, které jsou v daném okamžiku ve hře, uvědomujeme si pouze motivy dominantní, převládající.*
"Even with conscious motivation, we cannot be aware of all motives that play a role at a given moment; we are only realising the dominant, predominant motives."

Sentence (1) has two independent clauses, each defined by a main verb: *nemůžeme* "cannot" and *uvědomujeme* "realising". The first independent clause contains two clauses, as it has two verbs: *nemůžeme* "cannot" and *jsou* (translated here as "play"). The second independent clause consists of a single clause, as it contains only one finite verb.

The immediate unit of the clause is a phrase. A phrase is commonly understood as a multi-word constituent where one word, known as the head or root, holds prominence over the others (Osborne, 2019). For our purposes, the initial phrase of a clause includes the predicate and its leftmost dependent, along with all its own dependents. This approach emphasizes the linear progression of the sentence. Subsequent dependents of the predicate are treated as heads of their own phrases, each encompassing the head and all words dependent on it, directly or indirectly, if applicable.

A subphrase is a newly defined unit introduced by Nogolová et al. (2025). It is defined as the longest sequence of dependent words in which each word (except the head) has at most one dependent. Figure 2 provides a detailed view of the first clause in sentence (1). The left side of the figure illustrates the individual phrases. The first phrase includes the predicate and its leftmost dependent, while the second phrase consists of the remaining dependent – functioning as the head of the phrase – along with all of its own dependents. This results in the following division of the first clause:

[*Ani u uvědomované motivace nemůžeme*] [*mít přehled o všech motivech*].

The right side of the figure shows the individual subphrases. The first subphrase of the first phrase consists of the predicate *nemůžeme* on its own, because its dependent *u* has more than one dependent and cannot be included. The second subphrase is formed by *u* and all its dependents, since each of them has at most one dependent. Hence, the subphrases of the first phrase are [*nemůžeme*], [*Ani u uvědomované motivace*]. The second phrase makes up a single subphrase, as every word in it has at most one dependent.

A subphrase then consists of chunks defined according to the following criteria: (i) all words within the chunk share the same immediate parent (head); (ii) the chunk comprises only one level of dependency; (iii) the words are contiguous in the subphrase; (iv) no dependent word in the chunk has dependents outside of it.[5] Figure 3 presents the first phrase of the first clause of the sentence (1), divided into subphrases (on the left side) and individual chunks (on the right side). The first subphrase contains a single chunk, as it consists of only one word – *nemůžeme*. The second subphrase is divided into two chunks, [*Ani u*] and [*uvědomované motivace*], since only these combinations satisfy all the criteria outlined above.

Table 1 shows the number of sentences, independent clauses, clauses, phrases, subphrases, chunks, and words in language material we use (see Section 2). We include also words, as word length has been studied extensively in the last few decades. We thus can compare length distributions of syntactic units (a relatively new topic) with many previously achieved results for words.

Unit	**Tokens**
sentence	132 159
independent clause	167 132
clause	245 584
phrase	612 167
subphrase	697 244
chunk	1 047 142
word	1 535 506

Table 1: Number of units in the merged corpus.

Some of these units are, admittedly, purely formal, i.e., they are well defined, but, for the time

[4] All examples taken from Nogolová et al. (2025).

[5] This definition is taken from Anderson et al. (2019) with one modification; namely, Anderson et al. (2019) define chunks within sentences, while chunks as defined in this paper do not exceed the boundaries between subphrases.

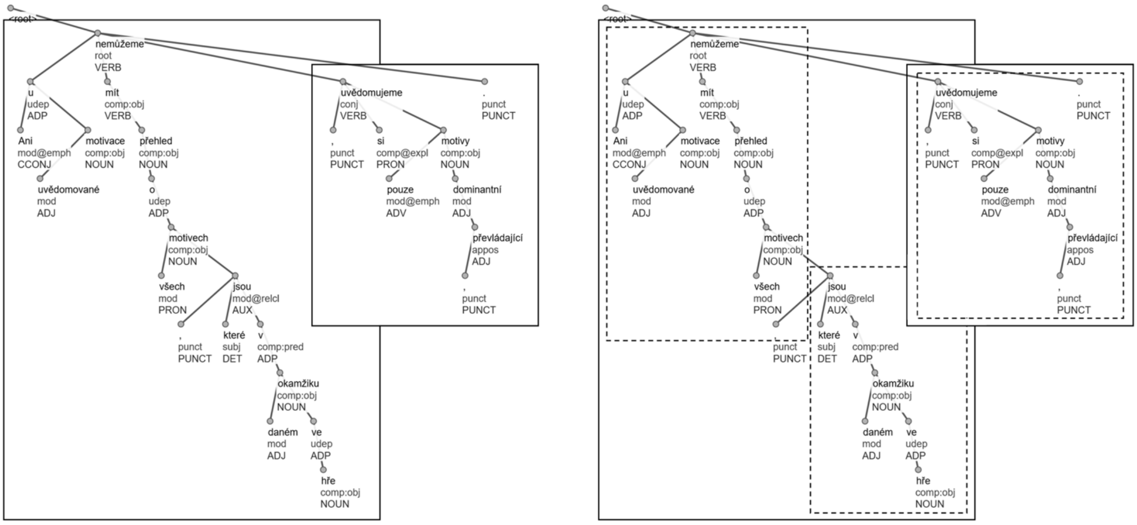

Figure 1: Dependency tree of sentence (1).

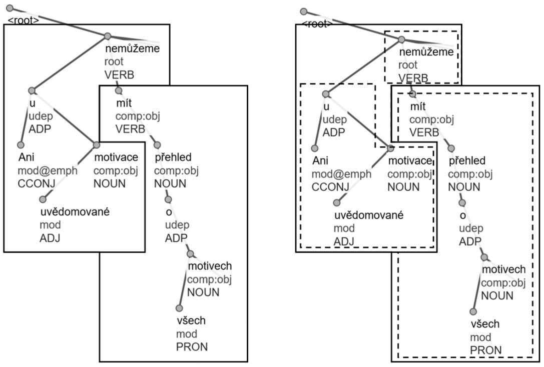

Figure 2: Phrase and subphrase structure of the first clause in sentence (1).

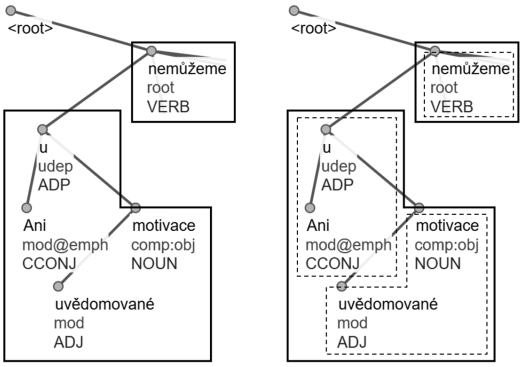

Figure 3: Chunks in both subphrases of the first phrase.

being, they do not have a linguistic interpretation. Nogolová et al. (2025) is a pilot study which shows that the MAL can indeed be modelled across many levels of the linguistic unit hierarchy. These units thus require investigations also from other points of view. This study focuses on modelling of their lengths.

4 Results

The hyper-Poisson distribution (Wimmer and Altmann, 1999, pp. 281-282) is defined as

$$P_x = \frac{a^x}{{}_1F_1(1;a;b)b^{(x)}}, \quad x = 0, 1, \dots \qquad (1)$$

with $a \geq 0$ and $b > 0$ being its parameters[6]; ${}_1F_1(1;a;b)$ is a hypergeometric function (see e.g. Gasper and Rahman, 1990). However, distribution (1) is defined on the set of all non-negative integers, while our data attain values from a finite set (no length exceeds the value of 13), and, moreover, it attains also the value of 0, while the lowest value of length in our data is 1. Therefore, we use a modification of distribution (1), namely, its right-truncated version shifted to the right by 1, with

$$P_x = C\frac{a^{x-1}}{b^{(x-1)}}, \quad x = 1, \dots, n, \qquad (2)$$

as a model for length distributions of the units from Section 3. The value of the parameter n is determined as the highest observed length value in the

[6] Symbol $b^{(x)}$ denotes the rising factorial, i.e., $b^{(x)} = \frac{\Gamma(b+x)}{\Gamma(b)}$.

data. For the normalization constant C it holds $C = \left(\sum_{x=1}^{n} \frac{a^{x-1}}{b^{(x-1)}}\right)^{-1}$, i.e., C is not an independent parameter, but its value depends on the values of the parameters $a \geq 0$ and $b > 0$; .

We express the goodness-of-fit of the model in terms of the determination coefficient R^2, with $R^2 \geq 0.9$ indicating a satisfactory fit (see Mačutek and Wimmer, 2013). We created a simple script in statistical software environment R[7] to fit this dataset. Estimated values of the parameters and the resulting determination coefficients are presented in Tables 2 and 3, and visualized in Figure 4.

One can see that the hyper-Poisson distribution fits all the data sufficiently well (all values of the determination coefficient are at least 0.9927). We remind the reader that the hyper-Poisson distribution is one of standard mathematical models for length of linguistic units. The Poisson distribution and its modifications and generalizations (including the hyper-Poisson distribution) and its applications to word length modelling can be found e.g. in Grzybek (2006) and Popescu et al. (2013). Thus, one can say that length of syntactic units (as they are defined in Section 3) behaves in the same way as word length. In addition, the hyper-Poisson distribution is a special case of a very general model of the linguistic theory from Wimmer and Altmann (2005). Therefore, although some of the new units used are waiting for their linguistic interpretation, they at least display a behaviour analogous to the units that are well established.

While the parameters of the hyper-Poisson distribution do not vary systematically between levels, there is a strong hint that their ratios b/a depend on the empirical repeat rate RR defined as $RR = \sum_{k=1}^{n} r_k^2$, with r_k being the relative frequency of length k, see Herdan (1962, pp. 36–40) and Altmann and Lehfeldt (1980, pp. 151–166). The repeat rate is a measure of diversity (or, from the opposite point of view, of uniformity) of the observed distribution. It can attain values from $1/n$ to 1. In the context of this paper, the value of $1/n$ corresponds to the same frequency of all lengths from 1 to n, whereas the value of 1 characterizes the deterministic distribution, i.e., all item have the same length. The values of the ratios b/a and the repeat rates RR for particular levels can be found in Table 4 and Figure 5; they are very strongly correlated, with the Pearson correlation coefficient of 0.973. This highly regular behaviour[8] opens a possibility of the interpretation of parameters in future research.

5 Conclusion

Analysis of the lengths of sentences, independent clauses, clauses, phrases, subphrases, chunks, and words revealed that these syntactic units exhibit length distribution patterns comparable to those found in more traditional linguistic units (especially in words). In each case, the length distribution can be modelled by the hyper-Poisson distribution, which has often been used as a mathematical model for (especially, but not only) word length. Moreover, based on Nogolová et al. (2025), these units consistently conform to the Menzerath–Altmann law across all structural levels, from sentences to syllables. Thus, they fit into the framework of interconnected linguistic units in which units and their properties are not isolated, but, rather, they influence each other (Köhler, 2005).

The results are promising, but we are aware of the fact that they are tentative only, as the only language from which we took data is Czech. Several important questions remain. These include the extent to which the findings can be generalized for other languages, as well as the potential influence of text genre and annotation schemes on the observed patterns. Further linguistic research will therefore be essential to determine the generality of the findings.

Acknowledgments

This research was supported by Moravian-Silesian Region, grant programme “Podpora vědy a výzkumu v Moravskoslezském kraji 2023”, project “Podpora talentovaných studentů doktorského studia na Ostravské univerzitě VII”, project No. 00516/2024/RRC (M. Nogolová), by projects APVV-21-0216 and VEGA 2/0120/24 (M. Koščová), by EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00748 (J. Mačutek), and by project “Lexikon a gramatika češtiny V - 2025 (MUNI/A/1445/2024)” (R. Čech).

[7] https://www.r-project.org/

[8] While the estimated parameter values depend also on the numerical estimation procedures, the ratio of b/a remains almost constant regardless of the method chosen.

length	**sentence in independent clauses**		**independent clause in clauses**		**clause in phrases**		**phrase in subphrases**	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	103 667	104 238.08	110 722	110 854.54	40 171	39 335.76	550 082	519 882.35
2	23 432	22 289.68	39 948	39 700.70	92 978	97 718.61	45 955	78 425.54
3	4 026	4 544.88	12 309	12 248.36	74 122	69 258.26	11 447	11 785.66
4	763	885.58	3 092	3 318.82	29 518	28 628.04	3 269	1 764.41
5	198	165.22	803	801.67	7 319	8 352.09	951	263.15
6	46	29.57	200	174.69	1 308	1 882.79	301	39.10
7	19	5.08	42	34.67	143	345.82	103	5.79
8	4	0.84	12	6.32	23	53.59	36	0.85
9	2	0.13	3	1.06	1	7.18	14	0.13
10	0	0.02	1	0.17	1	0.85	7	0.02
11	0	< 0.01					2	< 0.01
12	1	< 0.01						
13	1	< 0.01						
a	4.39		2.26		0.99		39.488	
b	20.53		6.21		0.40		261.72	
R^2	0.9998		> 0.9999		0.9954		0.9927	

Table 2: Fitting length frequencies by the hyper-Poisson distribution.

length	**subphrase in chunks**		**chunk in words**		**word in syllables**	
x	f_x	NP_x	f_x	NP_x	f_x	NP_x
1	438 584	444 369.74	591 655	591 539.21	471 587	461 580.80
2	183 484	167 238.90	396 848	392 045.32	487 216	514 703.06
3	48 494	58 448.27	51 432	58 364.50	345 045	333 384.51
4	17 746	19 066.26	6 495	4 894.08	168 721	152 164.10
5	5 936	5 831.10	649	285.64	49 843	53 616.00
6	1 994	1 678.51	58	12.768	10 544	15 384.27
7	658	456.34	5	0.46	2 065	3 723.02
8	215	117.54			362	778.99
9	65	28.76			96	143.56
10	34	6.70			19	23.643
11	5	1.49			4	3.52
12	1	0.32			3	0.48
13	1	0.06			1	0.06
a	4.90		0.19		1.55	
b	13.01		0.29		1.39	
R^2	0.9979		0.9998		0.9970	

Table 3: Fitting length frequencies by the hyper-Poisson distribution.

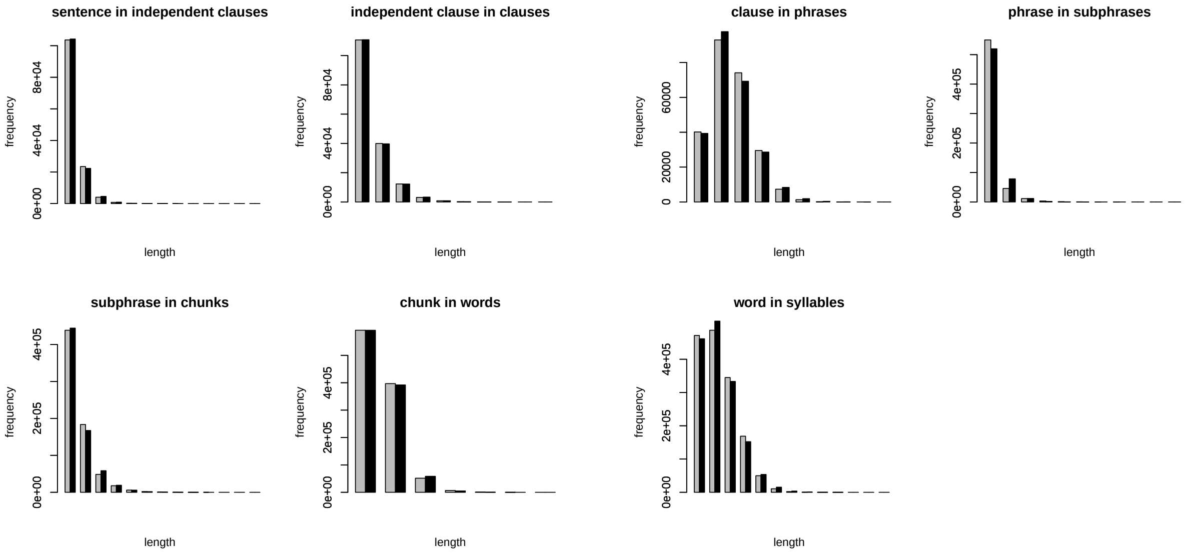

Figure 4: Fitting length frequencies by the hyper-Poisson distribution.

Unit	b/a	RR
sentence	4.677	0.648
independent clause	2.792	0.502
clause	0.403	0.277
phrase	6.629	0.813
subphrase	2.655	0.470
chunk	1.526	0.465
word	0.897	0.259

Table 4: Values of the repeat rate and of the ratio b/a.

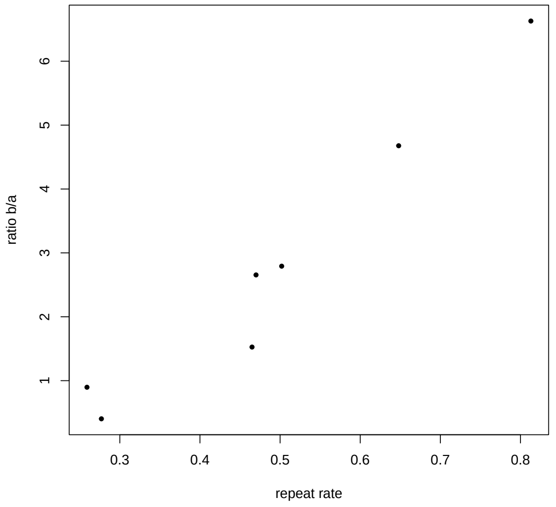

Figure 5: The relationship between the repeat rate and the ratio b/a.

References

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

Gabriel Altmann and Werner Lehfeldt. 1980. *Einführung in die Quantitative Phonologie*. Brockmeyer, Bochum.

Mark Anderson, David Vilares, and Carlos Gómez-Rodríguez. 2019. Artificially evolved chunks for morphosyntactic analysis. In Marie Candito, Kilian Evang, Stephan Oepen and Djamé Seddah (eds.), *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 133–143. ACL, Paris.

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

Eduard Bejček et al. 2013. Prague dependency treebank 3.0. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*, http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3.

Aleksandrs Berdicevskis. 2021. Successes and failures of Menzerath's law at the syntactic level. In Radek Čech and Xinying Chen (eds.), *Proceedings of the second workshop on quantitative syntax* (Quasy, SyntaxFest 2021), pages 17–32. ACL, Sofia.

George Gasper and Mizan Rahman. 1990. *Basic Hypergeometric Series*. Cambridge University Press, Cambridge.

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 (eds.), *Proceedings of the second workshop on universal dependencies* (UDW 2018), pages 66–74. ACL, Brussels.

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

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In Dimitra Gkatzia and Djamé Seddah (eds.), *Proceedings of the 16th conference of the European chapter of the Association for Computational Linguistics: System demonstrations*, pages 168–175. ACL, online.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. An introduction to word grammar. In Nancy Ide and James Pustejovsky (eds.), *Handbook of Linguistic Annotation*, pages 555–594. Springer, Dordrecht.

Gustav Herdan. 1962. *The Calculus of Linguistic Observations*. Mouton, The Hague.

Richard Hudson. 2010. *An Introduction to Word Grammar*. Cambridge University Press, Cambridge.

Tomáš Jelínek. 2017. FicTree: A manually annotated treebank of Czech fiction. In Jaroslava Hlaváčová (ed.), *ITAT 2017 Proceedings: Information technologies – applications and theory: Conference on theory and practice of information technologies*, pages 181–185. CreateSpace Independent Publishing Platform, Aachen, Technical University & Charleston.

Reinhard Köhler. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.), *Quantitative Linguistics. An International Handbook*, pages 760–774. de Gruyter, Berlin.

Vincent Kríž and Barbora Hladká. 2017. Czech Legal Text Treebank 2.0,. In *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*. Charles University, Prague. Http://hdl.handle.net/11234/1-2498.

Miroslav Kubát, Ján Mačutek, and Radek Čech. 2021. Communists spoke differently: An analysis of Czechoslovak and Czech annual presidential speeches. *Digital Scholarship in the Humanities*, 36(1):138–152.

Michal Křen et al. 2020. *SYN2020: reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague. Http://www.korpus.cz/.

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.

Ján Mačutek, Radek Čech, and Marine Courtin. 2021. The Menzerath-Altmann law in syntactic structure revisited: Combining linearity of language with dependency syntax. In Radek Čech and Xinying Chen (eds.), *Proceedings of the second workshop on quantitative syntax (Quasy, SyntaxFest 2021)*, pages 65–73. ACL, Sofia.

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

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. SUNY Press, Albany, NY.

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

Michaela Nogolová, Ján Mačutek, and Radek Čech. 2025. The Menzerath-Altmann law: From sentence to phoneme. *Journal of Quantitative Linguistics*, submitted paper.

Timothy Osborne. 2019. *A Dependency Grammar of English: An Introduction and beyond*. Benjamins, Amsterdam.

Petr Plecháč and Robert Kolár. 2015. The corpus of Czech verse. *Studia Metrica et Poetica*, 2(1):107–118.

Ioan-Iovitz Popescu, Sven Naumann, Emmerich Kelih, Andrij Rovenchak, Anja Overbeck, Haruko Sanada, Reginald Smith abnd Panchanan Mohanty, Andrew Wilson, and Gabriel Altmann. 2013. Word length: aspects and languages. In Reinhard Köhler and Gabriel Altmann (eds.), *Issues in Quantitative Linguistics 3*, pages 224–281. RAM-Verlag, Lüdenscheid.

Haruko Sanada. 2016. The Menzerath–Altmann law and sentence structure. *Journal of Quantitative Linguistics*, 23(3):256–277.

Milan Straka. 2018. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*, pages 197–207. ACL, Stroudsburg, PA.

Barbora Vidová Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. The Czech Academic Corpus 2.0 Guide. *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.

Gejza Wimmer and Gabriel Altmann. 1999. *Thesaurus of univariate discrete probability distributions*. Stamm, Essen.

Gejza Wimmer and Gabriel Altmann. 2005. Unified derivation of some linguistic laws. In Reinhard Köhler, Gabriel Altmann and Rajmund G. Piotrowski (eds.) *Quantitative Linguistics. An International Handbook*, pages 791–807. de Gruyter, Berlin.

Daniel Zeman et al. 2023. Universal Dependencies 2.13. *LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics*, https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5287.