Successes and failures of Menzerath's law at the syntactic level

Aleksandrs Berdicevskis
Språkbanken Text, Department of Swedish, University of Gothenburg
aleksandrs.berdicevskis@gu.se

Abstract

Menzerath's law is a quantitative generalization which predicts a negative correlation between the mean size of parts of a unit and the number of parts in the unit. In this paper, I use Universal Dependencies to perform a cross-linguistic test of Menzerath's law at two syntactic levels: whether the number of clauses in a sentence negatively correlates with mean clause length in this sentence and whether the number of words in a clause negatively correlates with mean word length in this clause. Menzerath's largely holds at the former level and largely does not at the latter. I discuss other interesting patterns observed in the data and propose some tentative partial explanations.

1 Introduction

Quantitative laws such as, for instance, Zipfian rank-frequency law (Piantadosi, 2014) or abbreviation law (Bentz and Ferrer-i-Cancho, 2016) are perhaps ones of the most universal generalizations that can be made about language. *Universal* here can be understood as both 'true for all / most languages' and 'true for various domains / levels of language'.

Another oft-cited generalization is Menzerath's law (Altmann, 1980; Stave et al., 2021), also called Menzerath-Altmann law. Menzerath's law predicts a negative correlation between the mean size of parts of a unit and the number of parts in the unit. Thus, the more sub-units (constituents) a linguistic unit (carrier unit, or construct) has, the shorter these units are expected to be on average. For instance, the more clauses a sentence contains, the shorter the mean length of these clauses (in words) is expected to be (Altmann, 1980).

Menzerath's law has been tested for various types of units in various languages (and also beyond language) and mostly (though not universally) found to be true (see an overview in Section 2). Most studies, however, used relatively small corpora (or even dictionaries), often of just one language, often not open-access, often shallowly annotated for the specific study. I use the Universal Dependencies (UD) collection to perform the largest-scale (to date) study of Menzerath's law at two syntactic levels: sentence–clause–word and clause–word–grapheme (see Section 3). I demonstrate that Menzerath's law works quite well at the former level, but not at the latter (see Sections 4, 5 and 6).

There is currently no consensus on *why* Menzerath's law emerges (or why it does not), and thus I cannot fully explain the observed results. In Section 7, however, I discuss which insights can be gleaned from the UD analysis and which hypotheses deserve further testing.

2 Background

2.1 Defining Menzerath's law

Any particular application of Menzerath's law has to be described at three levels: the length of a *unit* (for instance, clause), measured in *sub-units* (for instance, words), is supposed to negatively correlate with the mean length of sub-units, measured either in *sub-sub-units* (for instance, phonemes or graphemes) or

using a suitable continuous measure (for instance, seconds). In this paper, two triples will be analyzed: sentence–clause–word and clause–word–grapheme.

Menzerath's law has been shown to hold at different levels in different languages, but it is sometimes overlooked that there are at least two ways to interpret the claim *Menzerath's law holds*. One interpretation (which will be used in this paper) is 'the mean size of a sub-unit and the number of sub-units in the unit are negatively correlated' (Stave et al., 2021). Another interpretation is 'the relation between the mean size of a sub-unit (y) and the number of sub-units (x) can be approximated by a specific function'. The function is typically assumed to be $y(x) = ax^b e^{-cx}$ (Altmann, 1980), often simplified to $y(x) = ax^b$, though other variants have also been proposed (Milička, 2014). Sometimes the first interpretation is labelled as Menzerath's law, while the second one as Menzerath–Altmann's law (Ferrer-i-Cancho et al., 2014). Both interpretations rest on the assumption that number of sub-units and the mean size of sub-units are related (Mačutek et al., 2019).

While it is possible that there is a negative correlation *and* the relation can be approximated by Menzerath–Altmann's function (a power law with an exponential cutoff), it may also be that the latter is true, while the former is not. In Chinese, for instance, Menzerath–Altmann's function works well for sentence–clause–word ($R^2 = 0.85$) and clause–word–component ($R^2 = 0.77$; *component* is a constructing unit of a logogram) (Chen and Liu, 2019). Visual inspection of Chen and Liu's data, however, shows that the relation is clearly non-monotonic (down-up), and measuring Spearman's correlation coefficient shows there is no negative correlation for clause–word–component ($r = 0.42, p = 0.016$), while for sentence–clause–word the results are somewhat ambivalent ($r = -0.51, p = 0.052$). In a similar vein, Buk and Rovenchak (2008) report fitting Menzerath–Altmann's function for sentence–clause–word in Ukrainian, but the visualization of the data shows a non-monotonic (up-down) pattern, and Spearman's coefficient (calculated only for those sentence lengths which Buk and Rovenchak consider "reliable", that is, those for which at least 20 datapoints are available) does not show a negative correlation ($r = -0.13, p = 0.748$).

It can be argued that the latter, fit-a-model approach offers a more exact description of the reality. Following this logic, many studies (Cramer, 2005; Kelih, 2010; Baixeries et al., 2013; Milička, 2014) focus on finding the most appropriate formula and fine-tuning the parameters. On the other hand, if the model is complex enough, virtually any curve can be approximated reasonably well. To avoid overfitting, a clear theoretical explanation of the model is desirable. The existing explanations of Menzerath's law (see Section 2.3) mostly address the negative correlation, though attempts at explaining the Menzerath–Altmann's function and even interpreting its parameters have also been made (Köhler, 1984). I am not, however, aware of any explanation that would have successfully addressed the non-monotonic patterns observed above. Thus, in this study, Menzerath's law is understood as the negative correlation, without an attempt to describe the exact mathematical relation. The purpose of the study is to find out whether the law holds at the syntactic level.

2.2 Existing evidence

Altmann (1980, p. 129) predicts that Menzerath's law will hold for sentence–clause–word (otherwise the sentence presumably loses clarity). He also considers sentence–word–subword unit (word length can be measured in different ways: phonemes, graphemes, syllables, morphemes), but does not make a specific prediction, noting that "a monotonic decrease of word length can hardly be expected".

The first hypothesis (sentence–clause–word) has been tested before. Apart from the references mentioned in Section 2.1, Teupenhayn and Altmann (1984) report that Menzerath's law holds for German. Hou et al. (2017) find that in Chinese, it holds in formal written texts, but not in other registers. Xu and He (2020), however, demonstrate that in English, it holds for different registers. Roukk (2011), analyzing parallel texts in Russian and German and Russian and English, reports poor fitting results. Her data are too small for a correlation test to yield reliable results, but from a visual inspection it is obvious that there is no clear downward trend.

The second hypothesis (clause–word–subword unit) has received much less attention, but see the aforementioned study by Buk and Rovenchak (2008) and a relevant discussion by Altmann (1983)

Mačutek et al. (2017) look at clause–phrase–word in Czech, where *phrase* is defined as a subtree consisting of a node which is directly dependent of the clause predicate and all nodes that are (directly or indirectly) dependent on this node. They report good fitting results, and applying Spearman's test to their data (following their approach, only to those clause lengths that have more than 10 datapoints) yields a strong negative correlation ($r = -0.92, p = 0.001$).

Note that all these studies have at least one (often more) of the following limitations: they were performed for one language only; small corpora were used; those corpora had shallow annotation (for instance, number of clauses estimated by simply counting the number of finite verbs), often created specifically for the study; the data are not openly available.

In fact, the only large cross-linguistic study on Menzerath's law that I am aware of was performed by Stave et al. (2021), but it deals with the word–morpheme–grapheme level.

2.3 Explanations

There is no ubiquitously accepted explanation of why Menzerath's law is expected to hold. It has been argued to be mathematically trivial, but Ferrer-i-Cancho et al. (2014) provide evidence against this view.

It is typically assumed that Menzerath's law, similarly to Zipf's abbreviation law, emerges from efficiency pressures, but what exactly those pressures are is not fully understood. Köhler (1984) hypothesizes that the sub-units and the "structural information" about the connections between them must be stored at the same "register" in the brain (Vulanovic and Köhler, 2005). As the number of sub-units increases, so does the amount of structural information, and the only way to free up the necessary storage space is to use shorter sub-units. Milička (2014) develops this hypothesis further, but in both accounts the notion of structural information remains very vague.

Gustison et al. (2016) claim that Menzerath's law is caused by pressure for compression. They propose a unified formal mathematical framework for the explanation of Menzerath's law and Zipf's law of abbreviation.

It can actually be asked whether Menzerath's law cannot (at least in some cases) be reduced to Zipf's law of abbreviation. Consider, for instance, the word–morpheme–phoneme level. The more morphemes in a word, the higher the chance that many of them will be affixes rather than roots, have higher frequency and be on average shorter. Stave et al. (2021), however, show that for word–morpheme–grapheme, both Zipf's and Menzerath's law are at work, and removing one of them results in a poorer fit of a model.

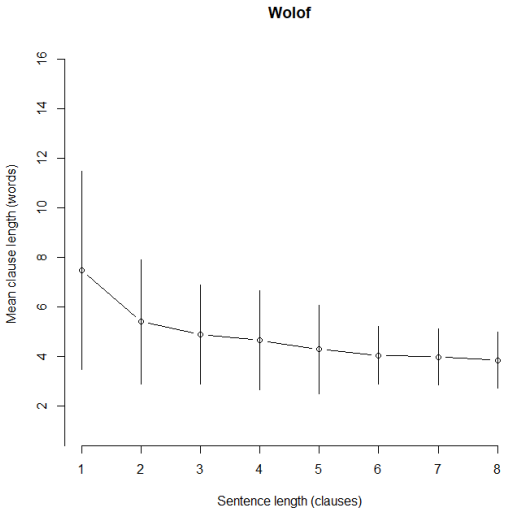
Coming back to syntax, the following level-specific explanation can be proposed for the sentence–clause–word level. Clauses often share certain elements. Open clausal component (raising and control structures, `xcomp` in UD), for instance, by definition does not have an internal subject, but the main clause may contain an element that functions as an (external) subject (cf. *Mary wants to buy a book*, where *Mary* is the subject of *wants*, but also the (external) subject of *buy*). Coordinated clauses can have shared arguments (*Mary is singing and dancing*, where *Mary* is the subject of both verbs), while repeated verbs can be omitted (gapping: *I like tea, and you coffee*). It can be expected that the number of clauses may correlate with the number of shared elements, thus reducing the average clause length. In a similar vein, clauses can act as elements of another clause. The length of the main clause *per se* can decrease, if dependent clauses fulfill the roles that would otherwise have been played by non-clausal dependents (see a test of this hypothesis in Section 4.2).

The present study is thus exploratory rather then confirmatory. It seeks to test whether Menzerath's law holds for sentence–clause–word and clause–word–grapheme across languages, and whether the cross-linguistic data lend support to any tentative explanations.

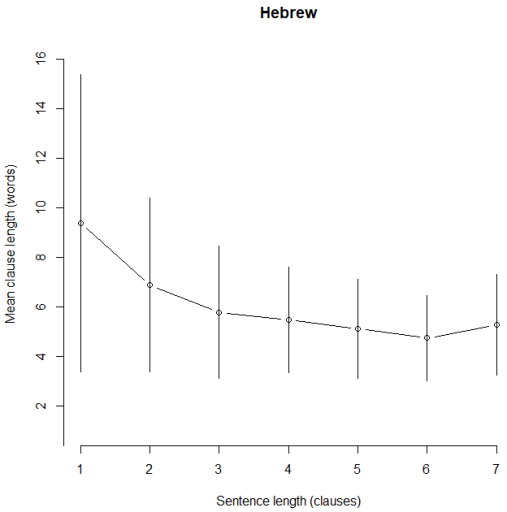
3 Materials and methods

I use corpus data from Universal Dependencies (UD) 2.8.1 (Zeman et al., 2021). All treebanks that do not have surface forms or that have less than 10,000 tokens are excluded from consideration. Naija-NSC treebank is also excluded, since it has an unusually high proportion (29%) of `dep` relation (which should normally be avoided).

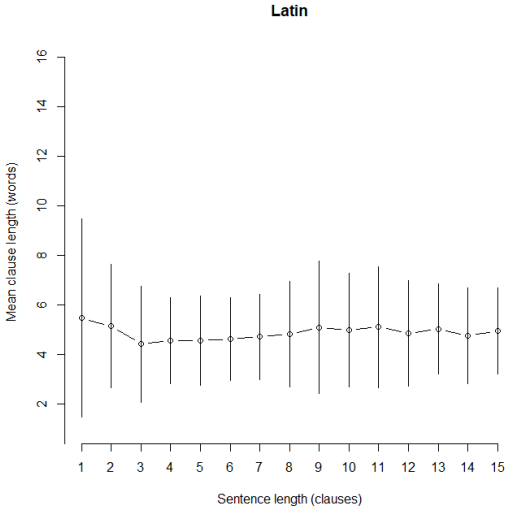
If a language has more than one treebank that fit the requirements, they are all concatenated. The



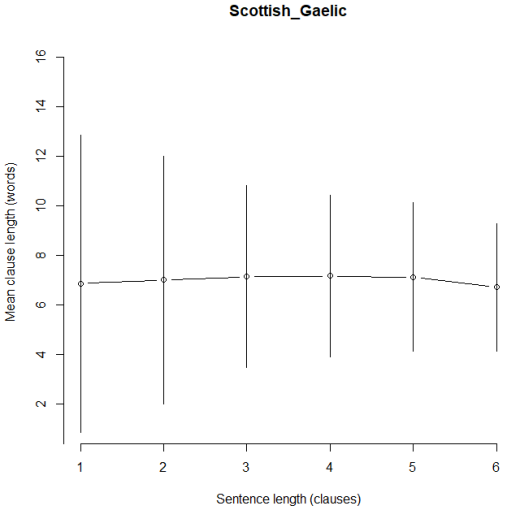
(a) Wolof. Perfect downward trend



(b) Hebrew. Downward trend with a deviation



(c) Latin. No clear downward trend



(d) Scottish Gaelic. No clear downward trend

Figure 1: Examples of correlation between sentence length (in clauses) and mean clause length (in words). Error bars show interquartile range. Sentence lengths with fewer than 50 datapoints were excluded

final dataset has 78 languages from 15 families (Indo-European, Afro-Asiatic, Mande, Basque, Mongolic, Sino-Tibetan, Uralic, Austronesian, Turkic, Mayan, Korean, Dravidian, Tai-Kadai, Austro-Asiatic, Niger-Congo). Note that how different genres are represented varies strongly across languages and treebanks. Genre is likely to affect the distribution of lengths of all units (sentences, clauses, words) and thus may potentially be a relevant factor. Nonetheless, since many treebanks do not have explicit detailed metadata about which sentence belongs to which genre, I do not attempt to control for genre.

The key notions ("sentence", "word", "clause") are operationalized as follows. "Sentences" are equivalent to UD sentences. Note, however, that sentence segmentation may not be a trivial task, for instance, for oral speech, ancient languages and social media, and thus even at the sentence level some inconsistencies across treebanks are possible.

"Words" are equivalent to UD tokens with minor exceptions. Punctuation marks (PUNCT) are excluded. Symbols (SYM) and unclassifiable tokens (X) must also be excluded (these labels can be used, for instance, for very long tokens like URLs, which can skew the results). However, unlike PUNCTs, SYMs and Xs can potentially have their own dependents and thus be important elements of the syntactic structure: it is not clear whether in such cases it is legitimate to remove them, but leave the rest of the sentence. For this reason, all sentences containing at least one SYM or X are excluded completely. Empty nodes (nodes with IDs like 1.1) are excluded, since they do not exist (should not inflate clause length) and do not have their own length. For multiword tokens, the token denoted by the range ID (e.g. 1–3), i.e. the surface token, is included in the analysis, the corresponding syntactic tokens (1, 2, 3) are excluded.

"Clauses" are most problematic, since there is no straightforward way to demarcate clauses in UD (as in most dependency grammars). Here, a clause consists of a node which has an incoming "clausal" relation (clausal root) and all descendants (both direct and indirect: children, grandchildren and so on) of the clausal root that do not have an incoming "clausal" relation.

Clausal relations are `root`, `csubj` (clausal subject), `ccomp` (clausal complement), `advcl` (adverbial clause modifier), `acl` (relative clause modifier), `parataxis`, some cases of `xcomp` (open clausal complement), some cases of `conj` (coordination). Relation subtypes are not distinguished (i.e. everything after a colon is ignored: `csubj:pass` is treated as `csubj`, `acl:relcl` as `acl`). `xcomp` is considered a clausal relation if its child is a verb, i.e. *great* in *You look great* does not start a new clause, while *work* in *I started to work there yesterday* does. `conj` is considered a clausal relation if either the parent or the child is a verb. The idea is to distinguish between clausal and non-clausal coordination, but the problem is that in cases of ellipsis, the head of a clause is not necessarily a verb. The rule "at least one of the conjuncts has to be a verb" covers some of such cases, but not the ones like *Jack a teacher, Jill a doctor*, which are possible and frequent in some languages.

Overall, the operationalization is necessarily crude and will certainly to some extent err on both sides: make false clause splits and fail to split when it should. Apart from the coordination problem, the following issues can be mentioned. Words (e.g. participles) can have `verb` as the POS label, but not actually behave like verbs. The `parataxis` relation can be argued to not always introduce a new clause. `dep`, `reparation` and `discourse` may deserve a special treatment, the former two should possibly be excluded, while the latter can be argued to introduce a new clause, at least in some cases. All these questions cannot be properly addressed without a thorough language- and treebank-specific linguistically-informed manual analysis.

Removing punctuation marks and empty tokens may in some cases lead to clauses or sentences consisting of zero words (e.g. if a clause/sentence consisted of an exclamation mark). All sentences where at least one clause has zero length are excluded.

Note that the language sample is not balanced (either genetically, areally or typologically). Excluding overrepresented language groups (mostly Indo-European) would lead to an undesirable data loss, since these languages also tend to have larger treebanks, which can be assumed to yield more robust and reliable results. To this end, I avoid averaging across languages (with the exception of clause type analysis in Section 4.2). Readers are encouraged to keep in mind that certain biases can emerge from the sample properties.

	negative	positive	none
sentence–clause–word	68	0	10
clause–word–grapheme	12	29	37
sentence–word–grapheme	26	19	30
sentence–clause–phrase	38	5	35
clause–phrase–word	58	2	18
phrase–word–grapheme	11	22	45

Table 1: Summary of the correlation tests across languages at different levels. The total amount of languages may vary across levels, since languages which do not have enough datapoints are excluded from the analysis

Section 4.2 describes an additional analysis that seeks to explore whether Menzerath's at the sentence–clause–word level can be explained by the fact that clauses share elements. Section 6 describes additional robustness analyses.

The code that was used to run the analyses and its detailed output are available at `https://github.com/AleksandrsBerdicevskis/menzerath`.

4 Results: Sentence–clause–word

4.1 General results

For every sentence in every language, I measured (according to operationalizations outlined in Section 3) how many clauses it contains and how many words the clauses in this sentence on average contain. I visually inspected the relation between the two variables for all languages. To prevent the results being skewed by outliers (usually a very small number of very long sentences), only those sentence lengths which had at least 50 datapoints were included. Note that languages vary greatly in terms of how many sentence lengths are represented in the data. After the 50-sentence filter is applied, Kiche, for instance, only has sentences which contain one or two clauses, while Icelandic covers the whole range from one to fifteen clauses.

Most typical patterns are represented by examples in Figure 1. In the vast majority of languages, the average clause length decreases monotonically according to what seems to be a power law (see, for instance, Wolof in Figure 1a). Sometimes, minor deviations from monotonicity are observed, often at large sentence length values (see Hebrew in Figure 1b). Nonetheless, even with the deviations most languages still exhibit a clear general downward trend. Those few for which the downward trend is not observed include, for instance, Latin (Figure 1c) and Scottish Gaelic (Figure 1d). In Latin, there is a decrease, but only in the beginning, while in Scottish Gaelic, there is rather a very small upward trend.

To do a formal test, I calculated Spearman's correlation coefficients between sentence length and clause length for all languages. They are reported in Table 3 in Appendix A (together with corresponding p-values) and summarized in Table 1. When summarizing the results, p-values, however, should be treated with caution, since they strongly depend on sample size, that is, how many different sentence lengths are represented in the data. For languages with small range of lengths p-values will never be small, even if perfect correlation is observed. Komi-Zyrian, for instance, demonstrates a perfect negative correlation, but only four different sentence lengths are represented (1–4 clauses), and the p-value is a theoretical minimum of 0.083. For this reason, the following criteria were applied. If the absolute value of the correlation coefficient was equal to or larger than 0.70, the language was labelled as demonstrating as either negative or positive correlation, regardless of the p-value. The same was done if the absolute value of the coefficient was larger than or equal to 0.30 and smaller than 0.70 and the p-value was smaller than or equal to 0.05. In other cases the correlation was assumed to be absent.

Under this interpretation, there are no cases of (anti-Menzerathian) positive correlation and 10 cases where the correlation could not be detected. It is difficult to tell whether it happens because it is truly absent or whether the sample is too small, and thus not clear whether the datapoints should be interpreted

as 'Menzerath's law does not hold' or 'Unknown whether Menzerath's law holds'. If we concentrate on languages with ranges 1–6 or larger (on the assumption that they yield more reliable samples), then seven languages out of 52 do not clearly conform to Menzerath's law: Latin, Scottish Gaelic, Icelandic, Old East Slavic, Old French, Finnish and Turkish. The other three ("small") languages that do not have a correlation are Manx, Breton and Sanskrit.

4.2 Is sharing elements across clauses the answer?

As a preliminary test of the hypothesis that Menzerath's law at the sentence–clause–word level can be explained by the fact that certain words are syntactically shared between clauses, I performed the following analyses.

First, I compared average lengths of various types of clauses. For every language, I extract all sentences that have exactly two clauses, one main (matrix) clause, one dependent (though in cases of coordination, the main–dependent contraposition is actually somewhat artificial). For dependent clauses, "type" is equivalent to the incoming relation of the clause root (that is, `ccomp`, `conj` etc.). For every clause type, its average length within language is calculated (types with less than 50 datapoints within a language were excluded). Note that if sentences which contain more than two clauses were included, the comparison would have to become much more nuanced. Main clauses could have different number of dependent clauses, while dependent clauses could have double roles: act as main clauses for their own dependents. These factors can potentially affect length distribution, and would have to be taken into account. For simplicity, the analysis is limited to two-clause sentences.

Clause lengths vary greatly across languages and treebanks. To correct for that and focus on the comparison across clause types within language, I normalized the average length of every type by average length of a simple sentence within the same language (that is, a sentence consisting of one and only one clause). These normalized lengths are then averaged across languages. The results are presented in Table 2. Keep in mind that such averaging may yield heavily skewed results, since the language sample is not balanced (and interquartile ranges suggest large variation for all types). Note also that not every clause type is represented in every language.

`xcomp`, as expected, tends to be short. The same is true for `parataxis`, probably because this relation is often used for short interjected clauses like parenthetical constructions (for example, *for example* or *of course*), tag questions etc. Interestingly, the longest type is not `main`, but `ccomp`.

As mentioned in Section 2.3, dependent clauses may perform functions of non-clausal dependents. `csubj` functions as a subject and is a clausal equivalent of `nsubj`, `advcl` is a clausal equivalent of `advmod`, `ccomp` and `xcomp` can be said to be clausal equivalents of `obj`, though note that this last correspondence is less clear. Consider now a main clause which has one of these clausal dependents, for instance, `csubj`. According to the operationalization used in this paper, the words contained by the dependent clause are not included into the main clause. In other words, there is a subject, but it is "outside" of a clause. If, however, the dependent was non-clausal (`nsubj`), it (and all its dependents) would have been inside the main clause and contributed to its length. It is no surprise then that main clauses are shorter than simple sentences. It is, however, interesting whether this is the only reason. To test that, I measure the decrease in length caused by having a clausal dependent (e.g. `csubj`) is approximately equal to the average length of a corresponding non-clausal dependent (`nsubj`).

I label every main clause in the two-clause-sentence sample described above by the type of dependent clause it has (`xcomp` and `ccomp` are merged together and labelled `comp`). The mean length of every "main-clause type" (normalized by the length of the simple sentence) across languages is reported in Table 2 in the column "main length".

Using the simple-sentence sample, I calculated the mean length (in words) of `nsubj`, `advmod` and `obj`. The column "diff" in Table 2 shows the normalized difference between the simple sentence length and the sum of two lengths: that of main-clause type (e.g. `csubj`) and the corresponding non-clausal dependent (`nsubj`). As all other numbers in the table, the difference is normalized by the simple-sentence length. If the hypothesis is correct, the difference should be close to zero, and indeed it is for `comp` and `advcl` (though note large interquartile ranges), but not for `csubj`.

Type	Length	IQR	Main length	IQR	Diff	IQR
ccomp	0.93	0.20	0.66	0.19	-	-
main	0.86	0.13	-	-	-	-
csubj	0.84	0.17	0.54	0.19	0.18	0.22
conj	0.78	0.17	0.90	0.16	-	-
advcl	0.74	0.16	0.89	0.17	-0.06	0.19
acl	0.72	0.19	1.06	0.21	-	-
xcomp	0.65	0.13	0.66	0.19	-	-
parataxis	0.62	0.24	0.82	0.19	-	-
comp	-	-	0.61	0.14	0.02	0.16

Table 2: Mean lengths across languages. The "main length" column should be read as 'mean length of a main clause having a dependent clause of the specified type'. "Diff" is a difference between the simple sentence length and the sum of two lengths: that of main-clause type ("main length") and the corresponding non-clausal dependent (e.g. textttnsubj for textttcsubj). All numbers are normalized by the mean length of a simple sentence in the same language. IQR = interquartile range.

No other clear patterns are observed. There does not seem to be any strong correlation between the length of the dependent clause of a certain type and corresponding main clause type. Interestingly, main clauses that have an `acl` clause are slightly longer than simple sentences.

5 Results: Clause–word–grapheme

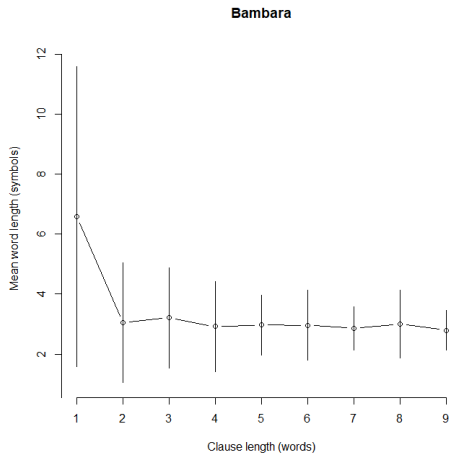
Exactly as with the sentence–clause–word analysis, I measured for every clause in every language how many words it contains and how many graphemes the words in this clause on average contain. I visually inspected the relation between the two variables for all languages. Again, only those lengths that had at least 50 datapoints were included.

Most typical patterns are represented by examples in Figure 2. Overall, the results were more variable than for the sentence–clause–word analysis, where one dominant pattern was observed. For clause–word–grapheme, 29 languages also exhibit a downward trend. Most often, it is L-shaped: a very steep decrease in the beginning, followed by a nearly flat line (see, for instance, Bambara in Figure 2a). In a few cases, the decrease is more gradually spread over the curve (Indonesian in Figure 2b). For 42 languages, an U-curve is observed, first a decrease and then a comparable increase (Latvian in Figure 2c). For four languages, the differences are so small that the pattern is best described as a flat line (see, for instance, Uyghur in Figure 2d). For four languages, there is an upward trend (Kazakh in Figure 2e). Finally, Persian (Figure 2f) exhibits a unique pattern: an inverted U-curve, an increase followed by a decrease.

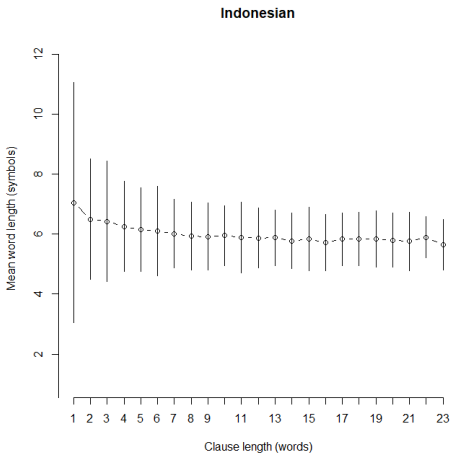
Spearman's correlation coefficients were calculated in the same way as for the sentence–clause–word level and summarized in Table 1. As can be seen from the summary, the adherence to Menzerath's law at the clause–word–grapheme level is much weaker. Note, however, that correlation coefficients are not really informative for the languages with clear non-monotonic patterns. Since I can propose no explicit hypothesis to explain the observed data, an inferential test is not appropriate: it is unclear *what* it can infer. Technically, some kind of non-linear regression model could of course be fitted to the data, but in the absence of a specific theory to test, the model would end up having many researcher degrees of freedom (Tong, 2019; Simmons et al., 2011), which is undesirable. I limit myself to labelling the observed curves as DOWN, UP, DOWN-UP, FLAT or UP-DOWN. The formalized procedure to determine the shape of the curve is described in Appendix B. The results are summarized above and reported in detail in Table 3 in Appendix A.

6 Robustness analyses

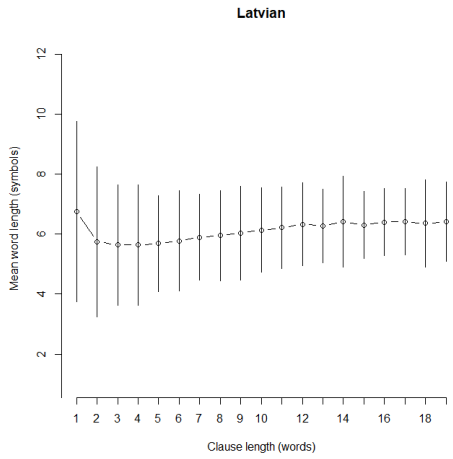
In order to test whether the results are robust, I reran the analyses with various thresholds instead of 50 datapoints per sentence/clause length (0, 20, 100). There were no qualitative changes of the overall



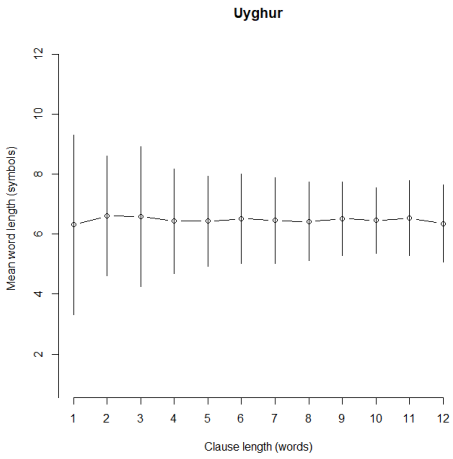
(a) Bambara. Downward trend (L-shape)



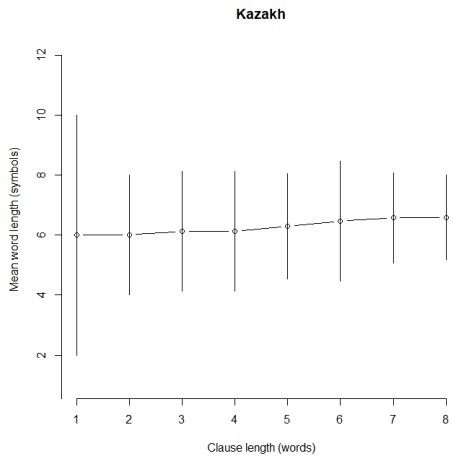
(b) Indonesian. Downward trend



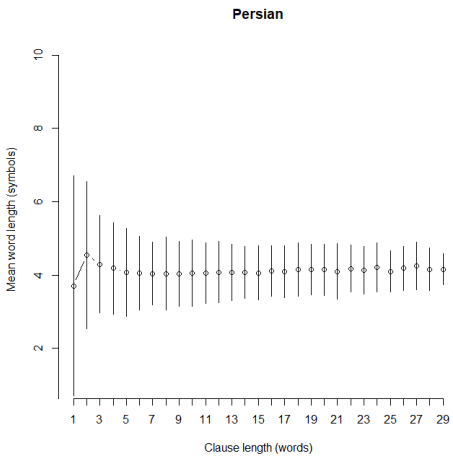
(c) Latvian. Down-and-up trend



(d) Uyghur. Flat line



(e) Kazakh. Upward trend



(f) Persian. Up-and-down trend

Figure 2: Examples of correlation between clause length (in words) and average word length (in graphemes). Error bars show interquartile range. Clause lengths with fewer than 50 datapoints were excluded

picture.

Since clause can be argued to be a problematic and / or imperfectly operationalized construct, I ran an analysis for the sentence–word–grapheme level (ignoring the clause level). The results resemble the ones for clause–word–grapheme (see Table 1).

Mačutek et al. (2017) reported that Menzerath's law holds for clause–phrase–word in Czech (see Section 2.2). It can be questioned whether their operationalization of phrase is theoretically adequate (in general, *phrase* is a less theory-neutral notion than *clause*), but I used it to run the analyses for clause–phrase–word and sentence–clause–phrase. I reproduced their findings for Czech, but overall, compared to sentence–clause–word, the adherence to Menzerath's law was slightly lower for clause–phrase–word, much lower for sentence–clause–phrase and even lower for phrase–word–grapheme (see Table 1).

7 Discussion

At the sentence–clause–word level, Menzerath's law largely holds, regardless of corpus size and typological or genealogical properties of language. It is not clear what is special about ten (or seven, depending on how one counts) languages that do not demonstrate an expected correlation. It can be noticed that four (or three) of them are ancient languages: Latin, Old East Slavic, Old French (and Sanskrit), but there are other ancient languages (e.g. Old Church Slavonic or Classical Chinese) that conform to Menzerath's law.

Clink and Lau (2020), analyzing primate communication, reach a somewhat similar conclusion: Menzerath's law holds in some cases, but not always (though in their study the adherence rate is much lower). They hypothesize that while the pressure for efficiency may facilitate compliance with Menzerath's law, other pressures may affect communication, sometimes to the extent that the law no longer holds. It is not, however, clear, which pressures could affect, for instance, non-Menzerathian Finnish and Icelandic, but not Menzerathian Estonian and Norwegian.

One potential confound is register, or genre (Hou et al., 2017). It is a question for future research to what extent Menzerath's law is robust to genre (and if it is not, why).

The explanation of why Menzerath's law (largely) holds is still wanting. The shared-element account that I propose seems to explain some cases, but not all, and it is not clear whether it is the sole reason. This hypothesis can potentially be further explored by using enhanced dependencies available in some UD treebanks (e.g. by measuring whether the shorter length of coordinated clauses can be "compensated" by taking into account shared dependents and elided verbs, or whether `xcomp` is shorter solely because it does not have an internal subject). Overall, it may be useful to consider whether Menzerath's law should be explained by level-specific factors, general optimization principles, or both.

At the clause–word–grapheme level, Menzerath's law generally does not hold. The observation by Altmann (1980) that the relationship is probably not monotonic turns out to be at least partly true. In the vast majority of cases, the mean word length as a function of number of words follows one of the two patterns: either L-shaped (steep decrease and then an almost flat line) or U-shaped (decrease and increase). L-shaped cases can be said to adhere to Menzerath's law, but first, they are less frequent than U-shaped ones, second, not all of them demonstrate a strong negative correlation.

Again, there does not seem to be any obvious way to explain the observed variance by different properties of languages or treebanks. Writing system may potentially be a confound. Apart from alphabets, the writing systems represented in the sample include (impure) abjads (vowels are omitted or partly omitted; e.g. Arabic), abugidas (consonant-vowel units are based on a consonant letter; vowel notation is secondary; e.g. Hindi) and logographic scripts (e.g. Mandarin). Japanese is a special case, using a mixture of a syllabary (kana) and a logographic script (kanji). However, if the writing system plays a role, its contribution is inconsistent: (Mandarin) Chinese and Classical Chinese do not conform to Menzerath's law, while Cantonese does; Hindi (Devanagari, abugida) does not, while Amharic (Ge'ez script, abugida) does.

It can be argued that graphemic word length is not the most adequate measure, and that phonemic length should be preferred. These measures, however, tend to be strongly correlated (Piantadosi et al., 2011). Moreover, should Menzerath's law hold for phonemes, it would probably mean that there is

some kind of optimization pressure due to which it emerges in oral speech. But then it is very likely that the same pressure would also affect written language (and most of the analyzed corpora contain predominantly written texts) and the law should hold for graphemes, too.

An anonymous reviewer raises two more important concerns. First, it can be questioned whether Menzerath's law should actually hold for *corpora* and not *texts* (cf. a discussion about inter- and intratextual laws by Grzybek and Stadlober (2011)). Given that the law is formulated as a relationship between the length of a unit and a sub-unit, and that it is hypothesized to emerge due to some kind of optimization pressure, I do not see any reason to assume that it should be valid only for texts and not for any sample of units, provided that the sample is large and representative. For any corpus, it can of course be questioned whether it is large and representative enough, but usually corpora tend to do better on these two scales than single texts.

The second concern is that Menzerath's law may not be valid if the unit, the sub-unit and the sub-sub-unit are not at the adjacent levels of the hierarchy. It can be argued that by testing the law on clause–word–grapheme, I am hopping over a level, since grapheme is not an immediate constituent of a word, and instead syllables or morphemes should be used. It is, however, unclear, which is the more appropriate unit, syllable or morpheme (or whether the law should work equally well for both). Furthermore, it is likely that graphemic (and phonemic) length is highly correlated with both syllabic and morphemic length. (To give an example: I measured the Spearman's correlation coefficient between the graphemic and the morphemic length of Swedish words, using the CoDeRooMor dataset (Volodina et al., 2021): $r = 0.83, p < 0.001$.) Note also that the robustness analyses described in Section 6 suggest that while adding or removing hierarchical levels (e.g. removing clause or adding phrase) affects the results, it does not change the overall picture. Nonetheless, this is a reasonable concern, and it would of course be beneficial to reproduce this study with syllable or morpheme as a sub-sub-unit. The problem is that the necessary resources are lacking.

Unlike Stave et al., I do not test for the role of Zipf's abbreviation law. For sentence–clause–word, this hardly is possible, since clauses are not repeated in languages often enough to enable frequency estimates. For clause–word–grapheme, I cannot propose an explicit prediction for the role of clause length that could have been tested by a regression model (see Section 5).

To conclude, Menzerath's law does not seem to be universal. It does not hold at some levels of analysis, and even at those where it does, some languages (or at least corpora) are exceptions. The reasons for that (both compliance and non-compliance) are not fully clear. Further studies should focus on explanatory approaches and on reproducing the existing results on larger and better samples.[1]

Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure *Nationella språkbanken*, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

References

Gabriel Altmann. 1980. Prolegomena to Menzerath's law. *Glottometrika*, 2:1–10.

Gabriel Altmann. 1983. H. Arens'«Verborgene Ordnung» und das Menzerathsche Gesetz. In Manfred Faust, Roland Harweg, Werner Lehfeldt, and Götz Wienold, editors, *Allgemeine Sprachwissenschaft, Sprachtypologie und Textlinguistik*, pages 31–39. Gunter Narr, Tübingen.

Jaume Baixeries, Antoni Hernández-Fernández, Núria Forns, and Ramon Ferrer-i-Cancho. 2013. The parameters of the Menzerath-Altmann law in genomes. *Journal of Quantitative Linguistics*, 20(2):94–104.

Chris Bentz and Ramon Ferrer-i-Cancho. 2016. Zipf's law of abbreviation as a language universal. In *Proceedings of the Leiden workshop on capturing phylogenetic algorithms for linguistics*, pages 1–4. University of Tübingen.

[1] Supplementary materials are available at `https://github.com/AleksandrsBerdicevskis/menzerath`.

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 *Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019)*, pages 25–32, Paris, France, August. Association for Computational Linguistics.

Dena J. Clink and Allison R. Lau. 2020. Adherence to Menzerath's law is the exception (not the rule) in three duetting primate species. *Royal Society Open Science*, 7(11):201557.

Irene Cramer. 2005. The parameters of the Altmann-Menzerath law. *Journal of Quantitative Linguistics*, 12(1):41–52.

Ramon Ferrer-i-Cancho, Antoni Hernández-Fernández, Jaume Baixeries, Łukasz Dębowski, and Ján Mačutek. 2014. When is Menzerath-Altmann law mathematically trivial? A new approach. *Statistical Applications in Genetics and Molecular Biology*, 13(6):633–644.

Peter Grzybek and Ernst Stadlober. 2011. Do we have problems with Arens' law? A new look at the sentence-word relation. In Peter Grzybek and Reinhard Köhler, editors, *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, pages 203–215. De Gruyter Mouton.

Morgan L. Gustison, Stuart Semple, Ramon Ferrer-i-Cancho, and Thore J. Bergman. 2016. Gelada vocal sequences follow Menzerath's linguistic law. *Proceedings of the National Academy of Sciences*, 113(19):E2750–E2758.

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.

Emmerich Kelih. 2010. Parameter interpretation of the Menzerath law: evidence from Serbian. In Peter Grzybek, Emmerich Kelih, and Ján Mačutek, editors, *Text and Language*, pages 71–80, Wien. Presens Verlag.

Reinhard Köhler. 1984. Zur Interpretation des Menzerathschen Gesetzes [On the interpretation of the Menzerath's law]. *Glottometrika*, 6:177–183.

Ján Mačutek, Radek Čech, and Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 100–107, Pisa, Italy, September. Linköping University Electronic Press.

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.

Jiří Milička. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics*, 21(2):85–99.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Maria Roukk. 2011. The Menzerath-Altmann law in translated texts as compared to the original texts. In Peter Grzybek and Reinhard Köhler, editors, *Exact Methods in the Study of Language and Text: Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*, pages 605–610. De Gruyter Mouton.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366. PMID: 22006061.

Matthew Stave, Ludger Paschen, François Pellegrino, and Frank Seifart. 2021. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard*, 7(s3):20190076.

Regina Teupenhayn and Gabriel Altmann. 1984. Clause length and Menzerath's law. *Glottometrika*, 6:127–138.

Christopher Tong. 2019. Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, 73(sup1):246–261.

Elena Volodina, Yousuf Ali Mohammed, and Therese Lindström Tiedemann. 2021. CoDeRooMor: A new dataset for non-inflectional morphology studies of Swedish. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 178–189, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Relja Vulanovic and Reinhard Köhler. 2005. Syntactic units and structures. In Reinhard Köhler, Gabriel Altmann, and Rajmund Piotrowski, editors, *Quantitative linguistics: An international handbook*, pages 274–291. Walter de Gruyter, Berlin.

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.

Daniel Zeman, Joakim Nivre, et al. 2021. 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.

Appendix A. Detailed results across languages

	sentence–clause–word			clause–word-grapheme			general info	
language	*r*	*p*	range	trend	min	range	corpus size	family
Afrikaans	-1.00	0.083	4	down	4	18	49	ine
Akkadian	-0.80	0.333	4	down*	10	11	25	afa
Amharic	-1.00	0.333	3	down*	5	5	10	afa
Ancient Greek	-0.75	0.018	10	down*	15	22	417	ine
Arabic	-0.96	0.003	7	up*	5	22	303	afa
Armenian	-0.96	0.003	7	down-up	3	14	53	ine
Bambara	-0.90	0.083	5	down*	9	9	14	dmn
Basque	-1.00	<0.001	7	down-up	6	15	121	eus
Belarusian	-0.96	0.003	7	down-up	5	21	305	ine
Breton	-0.50	1.000	3	down	5	9	10	ine
Bulgarian	-1.00	0.003	6	down-up	3	22	156	ine
Buryat	-1.00	0.083	4	down-up	7	9	10	xgn
Cantonese	-0.90	0.083	5	down-up	2	10	14	sit
Catalan	-1.00	<0.001	9	down	8	36	547	ine
Chinese	-0.96	<0.001	11	flat	1	20	285	sit
Clas. Chinese	-1.00	0.017	5	up*	1	13	269	sit
Coptic	-1.00	<0.001	7	down	4	8	49	afa
Croatian	-1.00	<0.001	8	down-up	4	23	199	ine
Czech	-0.99	<0.001	10	down-up	3	36	2223	ine
Danish	-1.00	0.003	6	down-up	5	20	101	ine
Dutch	-1.00	0.003	6	down-up	5	25	307	ine
English	-0.97	<0.001	9	down-up	3	26	556	ine
Erzya	-1.00	0.017	5	down*	6	8	17	urj
Estonian	-0.93	0.001	9	down-up	4	19	507	urj
Faroese	-0.79	0.048	7	down	4	13	50	ine
Finnish	-0.94	0.017	6	down-up	4	15	397	urj
French	-0.98	<0.001	9	down	7	32	583	ine
Galician	-1.00	0.003	6	down	8	28	164	ine
German	-1.00	<0.001	7	down	2	35	3754	ine
Gothic	-0.94	0.017	6	down-up	8	14	55	ine
Greek	-0.94	0.017	6	down	13	19	63	ine
Hebrew	-0.89	0.012	7	down	7	19	161	afa
Hindi	-1.00	0.017	5	down-up	6	34	376	ine
Hungarian	-1.00	0.017	5	down-up	4	19	42	urj
Icelandic	-0.06	0.822	15	down-up	3	28	1162	ine
Indonesian	-0.89	0.012	7	down*	23	23	168	map
Irish	-0.94	0.017	6	down-up	3	23	131	ine
Italian	-1.00	<0.001	9	down-up	5	34	819	ine
Japanese	-0.98	<0.001	9	down	20	25	237	jpx
Kazakh	-1.00	0.333	3	up*	1	8	11	trk
Kiche	-1.00	1.000	2	down*	7	8	10	myn
Komi Zyrian	-1.00	0.083	4	down*	8	8	10	urj
Korean	-0.99	<0.001	10	down*	16	17	447	(kor)
Kurmanji	-1.00	0.333	3	down-up	5	11	10	ine
Latin	0.12	0.676	15	down-up	15	31	978	ine
Latvian	-1.00	<0.001	8	down-up	4	19	252	ine

Continued on next page

Table 3 – continued from previous page

	sentence–clause–word			clause–word-grapheme			general info	
language	*r*	*p*	range	trend	min	range	corpus size	family
Lithuanian	-1.00	<0.001	7	down-up	5	14	75	ine
Maltese	-1.00	<0.001	8	down	6	15	44	afa
Manx	0.50	1.000	3	down-up	5	10	21	ine
North Sami	-0.80	0.333	4	down-up	4	10	27	urj
Norwegian	-0.93	0.002	8	down-up	3	26	667	ine
OCS	-0.82	0.034	7	down*	10	13	58	ine
OES	-0.18	0.713	7	down-up	6	20	180	ine
Old French	-0.29	0.556	7	down-up	8	17	171	ine
Persian	-0.88	0.003	9	up-down	1	29	655	ine
Polish	-1.00	<0.001	9	down-up	4	22	499	ine
Portuguese	-0.93	0.001	9	down	21	34	571	ine
Romanian	-0.86	0.001	11	down-up	5	32	938	ine
Russian	-0.87	0.001	11	down-up	5	28	1421	ine
Sanskrit	-0.70	0.233	5	down	8	10	29	ine
Scottish Gaelic	-0.03	1.000	6	down-up	3	19	72	ine
Serbian	-1.00	<0.001	7	down-up	3	20	98	ine
Slovak	-0.90	0.083	5	down-up	4	16	106	ine
Slovenian	-1.00	<0.001	7	down-up	3	20	170	ine
Spanish	-0.99	<0.001	11	down	22	37	1015	ine
Swedish	-1.00	0.083	4	down-up	4	14	207	ine
Tamil	-1.00	0.083	4	down	5	10	12	dra
Thai	-1.00	<0.001	7	down-up	4	14	22	(tai)
Turkish	-0.45	0.267	8	down	4	20	592	trk
Turkish German	-1.00	0.003	6	down*	14	14	37	ine
Ukrainian	-1.00	<0.001	7	down-up	3	19	122	ine
Upper Sorbian	-1.00	0.333	3	up	5	11	11	ine
Urdu	-1.00	0.017	5	down-up	6	30	138	ine
Uyghur	-1.00	0.017	5	flat	1	12	40	trk
Vietnamese	-1.00	0.017	5	down	7	8	44	aav
Welsh	-1.00	0.017	5	down-up	6	17	37	ine
West. Armenian	-0.86	0.024	7	down-up	6	14	36	ine
Wolof	-1.00	<0.001	8	down-up	2	13	44	nic

Table 3: Results across **languages** (OCS = Old Church Slavonic, OES = Old East Slavic). For **sentence–clause–word** analysis: *r* = Spearman's correlation coefficient, *p* = corresponding *p*-value, range = maximum sentence length (in clauses) for which 50 datapoints are available. For **clause-word-grapheme** analysis: trend = the shape of the curve, min = clause length for which the shortest mean word length is observed, range = maximum clause length (in words) for which 50 datapoints are available. For languages with DOWN, UP or FLAT trend, asterisk marks those where $|r| \geq 0.70$ or $|r| \geq 0.30$ and $p \leq 0.05$. Corpus size is given in K words, families are denoted by ISO-639 codes. There are no ISO-639 codes for Koreanic (the code for Korean is used) and Tai-Kadai (the code for the Tai branch is used).

Appendix B. The procedure for determining the shape of the curve

The formal procedure of determining the shape of the curve ("trend") for the clause–word–grapheme (reported in Table 3 in Appendix A) was as follows. The extrema (maximum and minimum) of the curve were identified. Then four points (first point, the smallest clause length; maximum; minimum; last point,

the largest clause length) were compared by means of *t*-tests between adjacent pairs of points. In many cases, there were actually only three or even two points, because either maximum or minimum (or both) coincided with either first or last point (or both). Thus, the number of *t*-tests varied from one to three (Bonferroni correction for multiple comparisons was applied). If *p*-value was smaller than 0.05 and the absolute value of Cohen's *d* (effect size) was larger than 0.20, then the difference was considered to be large enough to label the corresponding part of the curve as going either DOWN or UP, otherwise it was ignored. If there were no differences at all, the whole curve was labelled as FLAT.

Bear in mind that the procedure is descriptive rather than inferential (even though it uses inferential statistics as a technique). It is approximately equivalent to manually classifying the patterns, but relies on formalized criteria and thus is more reproducible. See main text for the reasons why more sophisticated inferential tests were not applied.