Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws

Matthew Stave, Ludger Paschen, François Pellegrino, Frank Seifart

Matthew Stave*, Ludger Paschen, François Pellegrino and Frank Seifart

Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws

https://doi.org/10.1515/lingvan-2019-0076
Received November 1, 2019; accepted October 23, 2020

Abstract: Zipf's Law of Abbreviation and Menzerath's Law both make predictions about the length of linguistic units, based on corpus frequency and the length of the carrier unit. Each contributes to the efficiency of languages: for Zipf, units are more likely to be reduced when they are highly predictable, due to their frequency; for Menzerath, units are more likely to be reduced when there are more sub-units to contribute to the structural information of the carrier unit. However, it remains unclear how the two laws work together in determining unit length at a given level of linguistic structure. We examine this question regarding the length of morphemes in spoken corpora of nine typologically diverse languages drawn from the DoReCo corpus, showing that Zipf's Law is a stronger predictor, but that the two laws interact with one another. We also explore how this is affected by specific typological characteristics, such as morphological complexity.

Keywords: corpus linguistics; cross-linguistic; efficiency; language universals; Menzerath; typology; Zipf

1 Introduction

Two principles that have been postulated to predict the length of linguistic elements, e.g., word length in terms of number of phonemes, are Zipf's Law of Abbreviation (Zipf 1935, 1949) and Menzerath's Law (Menzerath 1928). Zipf's Law describes a negative correlation between the length of an element and its text frequency. Menzerath's Law describes a negative correlation between the length of a carrier unit and the lengths of its sub-units, e.g. the length of a word and its component syllables. Both of these laws have been demonstrated to affect not only lengths of elements (measured in phonemes or graphemes) but also durations (measured in milliseconds). The analyses below, based on corpus data, will deal exclusively with graphemic length.

Zipf's Law was originally demonstrated with graphemic word length in text corpora (Zipf 1949), showing that more frequent words tend to be shorter, and has proven cross-linguistically robust across nearly 1,000 languages (Bentz and Ferrer-i-Cancho 2016). It has also been documented above the word level, in n-gram sequences for English (Smith and Devine 1985). Zipfian frequency is also hypothesized to correlate with the presence or absence of morphological marking – and in some cases, length of morphological markers, from a typological perspective (Haspelmath 2018). Regarding word length, it has been shown that word length correlates more closely with contextual predictability, rather than raw frequency (although both are closely related) (Piantadosi et al. 2011). The same principle also manifests in temporal compression of more frequent elements, both for humans (Strunk et al. 2019) and for dolphins and Formosan macaques (Ferrer-i-Cancho et al. 2013).

In his original formulation, Zipf explained the Law of Abbreviation by the Principle of Least Effort, under which speakers minimize the effort in production by producing shorter words proportionally to how often they

***Corresponding author: Matthew Stave**, Univ Lyon and Centre National de la Recherche Scientifique (UMR 5596), Dynamique du Langage, Lyon, France, E-mail: stave.matthew@gmail.com. https://orcid.org/0000-0002-8590-8856
Ludger Paschen and Frank Seifart, Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany. https://orcid.org/0000-0001-9909-2088 (F. Seifart)
François Pellegrino, Univ Lyon and Centre National de la Recherche Scientifique (UMR 5596), Dynamique du Langage, Lyon, France

must produce them. The Principle of Least effort is counter-balanced by a similar principle of effort reduction, but for the listener. Speakers try to minimize the amount of effort that must go into comprehending what they say by making the linguistic production as clear and unambiguous as possible. The Law of Abbreviation predicts that the Principle of Least Effort will win out in highly frequent words, maximizing the number of tokens that are reduced, and minimizing the number of types. Note that it is still a controversial issue, whether Zipf's law results from social-cognitive optimization processes or more general underlying principles at play in most natural and cultural systems (e.g. Mandelbrot 1953; Miller 1957).

Menzerath's Law was originally demonstrated with graphemic syllables in German words (Altmann 1980; Menzerath 1928, see also Fenk and Fenk-Oczlon 1993), showing that the longer the word, the shorter the mean length of the syllables in the word. Since then, it has also been described for clause lengths in multi-clausal sentences (Teupenhayn and Altmann 1984), word lengths in sentences and clauses (Köhler 1982), and arguments of verbs (Mačutek et al. 2017). Like Zipf's law, it extends to temporal compression, as evidenced in polysyllabic shortening in human language (Lehiste 1970), and in durations of chimpanzees' vocal calls as a function of the number of calls in a sequence (Fedurek et al. 2017).

Altmann (1980) gave a mathematical formulation to Menzerath's Law, but the underlying motivation for this law is less intuitively clear than that of Zipf's Law. One explanation given relates to two different kinds of information encoding (Köhler 1984; Milička 2014). The first is semantic, grammatical, or phonological information that the sub-element contributes on its own, called 'plain information' by Köhler and Milička. This kind of information is assumed to increase with the length of the sub-element. For instance, the longer a sub-element like a clause or word is, the more information it will carry. The second is information about other sub-elements (i.e. contextual information about other sub-elements and the structural makeup of the carrier element), called 'structural information' by Köhler and Milička. This kind of information is assumed to increase with the number of sub-elements. For example, within a complex clause, a relative clause may not need to express a subject because of its structural relationship to the matrix clause, or the presence of a preposition may signal information about the function of a neighboring noun phrase. As the number of sub-elements increases, so does the amount of structural information, thus the amount of plain information can decrease, leading to a decrease in the length of the sub-elements. An alternative way to explain this law is in terms of uncertainty reduction. From this perspective, an element's plain information is the uncertainty reduction due to the occurrence of the sub-element, and the structural information is the uncertainty reduction due to the co-occurrence of the sub-elements once the plain information of each individual sub-element is accounted for.

These two laws can be seen as contributing to language efficiency by regulating the length of linguistic elements in terms of an optimization of cost/benefit ratio of effort in speech production to successful transmission. But while both laws explain the benefit of reduced production effort, the trade-off in terms of cost appears to be different for each: for Zipf's Law, the cost is decreased comprehensibility, while for Menzerath's Law the cost is better explained in terms of greater effort in working memory and planning. Given these differences, it is not clear how the two laws interact to predict lengths of linguistic units in the world's languages.

We will explore the contributions of these laws in a cross-linguistic analysis of morpheme lengths in nine typologically diverse languages. Morpheme lengths have not been widely studied so far: Zipf found that morphemes obey the Law of Abbreviation in German (Zipf 1935), and Menzerath's Law has been shown to apply to morpheme length in languages such as German (Gerlach 1982) and Czech (Milička 2014), but morpheme lengths have not received a systematic cross-linguistic analysis, due to the fact that morphologically-annotated corpora are time-consuming to create. These two laws have also rarely been studied in parallel, though see Heesen et al. (2019) for an examination of the two laws in animal communication.

In our analyses, we ask to what extent Zipf's and Menzerath's Laws influence the length of morphemes, and whether this is constrained by specific typological characteristics. In a pair of regression analyses, we examine whether one law is more predictive of morpheme length than the other, and whether the two laws function independently of each other. We then explore how these laws vary across languages to assess typological effects.

2 Materials and methods

The corpora used in this analysis are drawn from the DoReCo (Language DOcumentation REference COrpus) database (Paschen et al. 2020). DoReCo is a collection of 50+ spoken corpora of mostly lesser-documented languages, each with a minimum of 10,000 words. They are time-aligned at the phoneme level using the MAUS time-alignment software (Kisler et al. 2017), and will be publicly available in 2021.[1] A subset of 30 corpora are morphologically annotated, with morpheme breaks, glosses, and part of speech tags. The database is designed to be a typologically and areally diverse convenience sample. At this stage of the project a sample of nine corpora is sufficiently processed and will be used in the current study. The morphological annotations have been provided by field linguists who are experts in the language. The texts are primarily personal and traditional narratives.

Information on the nine languages examined below is summarized in Table 1. They represent eight language genera (Dryer 1989) from seven macro-families. They represent a diverse sample of morphological complexity, from mostly isolating to fairly synthetic languages. The synthesis index, which measures the average number of morphemes per word, was calculated from the corpora, and ranges from 1.10 (Fanbyak) to 2.71 (Hoocąk). This captures a large portion of the range of morphological complexity in human languages, especially regarding isolating languages, which reach extremes such as Vietnamese (1.06), although it lacks languages at the upper end, which feature extremes like West Greenlandic (3.72) (Haspelmath and Sims 2013: 6).[2]

Table 1: Overview of the language sample, with Glottolog language identification codes (Hammarström et al. 2020), and information on corpora used in these analyses.

Language	Family	Genus	Morphology	Synthesis index	Corpus size (wds)	Glotto-code	Reference
Fanbyak	Austronesian	Malayo-Polynesian	Mostly isolating	1.10	14,388	orko1234	Franjieh (2018)
Goemai	Afro-Asiatic	West Chadic	Mostly isolating	1.32	45,680	goem1240	Hellwig (2003)
Kakabe	Mande	Western Mande	Verbal and nominal suffixation	1.54	45,127	kaka1265	Vydrina (2013)
Sumi	Sino-Tibetan	Kuki-Chin	Some prefixation and compounding	1.65	23,740	sumi1235	Teo (2013)
Totoli	Austronesian	Malayo-Polynesian	Synthetic	2.05	12,997	toto1304	Leto et al. (2010)
Katla	Atlantic-Congo	Katla-Tima	Synthetic - agglutinative	2.12	17,071	katl1237	Hellwig (2007)
Urum	Turkic	Common Turkic	Synthetic - agglutinative	2.16	20,773	urum1249	Skopeteas and Moisidi (2011)
Gorwaa	Afro-Asiatic	Southern Cushitic	Synthetic - fusional	2.23	14,012	goro1270	Harvey (2016)
Hoocąk	Siouan	Core Siouan	Synthetic	2.71	22,515	hoch1243	Hartmann (2004)

Our units of analysis are segmental morphemes, or, more specifically, morphs, as represented orthographically in the corpus, without distinguishing between lexical and grammatical morphemes. The current study also does not consider the number of meanings per morph, as in fusional morphemes or additional suprasegmental information, though from an information-theoretic perspective these meanings likely play an important role. In these corpora, we consider words to be space-separated character strings; clitics are typically coded as affixes, so the word units approximate prosodic words.

1 See http://doreco.info for more information on the project.

2 For one analysis of morpheme length in a polysynthetic language (Lakota), see Pustet and Altmann (2005), which finds an interesting multi-modal frequency distribution of morpheme lengths, which they attribute to syllable structure constraints (although this paper does not consider them in the context of Menzerath's Law).

To assess the relative strengths of Zipf's Law of Abbreviation and Menzerath's Law, we took measures from the nine corpora.[3] As a dependent variable, we use the grapheme length of each morpheme (morph_len). Grapheme length does not correspond perfectly to phoneme length or articulatory effort, but has been widely used in studies of both Zipf's and Menzerath's Laws as a reliable proxy for effort in production. This is justified by the fact that the correlations between grapheme length and phoneme length are very high, even for languages with relatively deep orthographies like English and Dutch (Piantadosi et al. 2011).

For Zipf's Law, which predicts element length based on element frequency, we took the log token frequencies of all morphemes, normalized for each corpus (norm_freq). For Menzerath's Law, which predicts morpheme length from the length of the carrier word, we took the mean word length of all words that carry the morpheme in question, to represent the overall effect of the length of the carrier unit on the morpheme, using word types instead of word tokens. Word types capture the range of embeddings a morpheme has, and give a picture of how the morpheme can be integrated into word-forms. Word tokens capture the degree of activation a given morpheme has in a particular word, biasing the picture towards these frequent word-forms. Token frequency also biases the sample towards mono-morphemic words, which are more frequent than multi-morphemic words, and which offer less scope for Menzerath's Law to have an effect. Adding to this, Menzerath's Law is expected to be due to an intrinsic trade-off between the components and the carrier, and not to the frequency of the of usage of the specific carrier. For these reasons, we use word types in the analyses below, rather than word tokens.

We determined two ways of assessing the strength of the carrier word's length: word length in graphemes (word_len) and word length in number of morphemes (morph_num). Measuring the grapheme length of the word targets, to some extent, the effort in articulation, while measuring the length in morphemes targets the cognitive effort in morphological processing. Both measures are included in the analyses.

To get a picture of the structure of the data used here in terms of morpheme and word lengths, Figure 1 plots the mean morpheme length and mean word length of the languages in the sample. Mean word lengths are mostly between 5 and 8, and mean morpheme lengths are mostly between 4.5 and 6.5. The exception is Hoocąk, which has a mean word length of 13 and a mean morpheme length of nearly 7. Hoocąk is the most synthetic language in the sample, and employs frequent bi-graphemic long vowels, which explain its position in Figure 2. As expected, there is a positive correlation between word length and morpheme length, with more synthetic languages clustering towards the longer edge of the distribution.

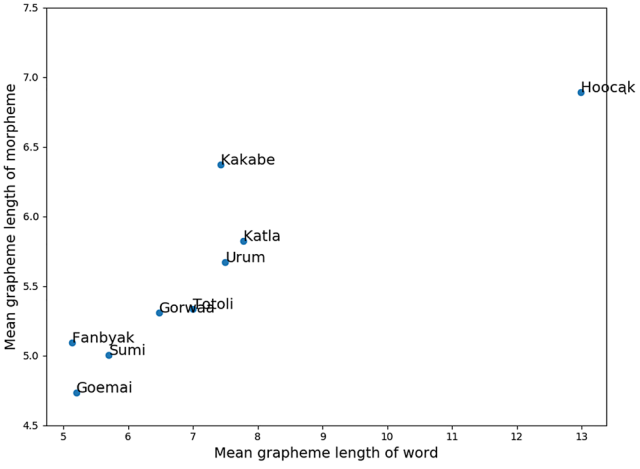

Figure 1: Mean grapheme length of words and morphemes by language.

3 All analyses were done in Python, using the pandas, math, and statsmodels packages.

Figure 1 demonstrates that morpheme length is highly collinear with word length (because words contain morphemes), so instead of using the total word length as an independent variable, we will use the mean length of the carrier words minus the length of the morpheme: essentially, the mean remaining length of the word after the morpheme is removed.

Ordinary least squares (OLS) linear regression analyses were then run, using morpheme length as the dependent variable (no zero morphemes are used, so all morphemes have a length of at least 1). The three independent variables – normalized log frequency of morpheme (norm_freq), mean length of carrier words minus morpheme length (word_len), and mean number of morphemes in carrier words (num_morph) – were all z-scored, and outliers beyond three standard deviations were removed from the analysis, resulting in an omission of less than 5% of the total data. After this, we examined the language-specific correlations between morpheme length and each of the independent variables, to examine the role of specific typological characteristics.

3 Results

The Zipfian log frequencies are shown below in Figure 2. Each language shows clear tendencies towards the expected distribution of higher frequencies for shorter morphemes, with some exceptions, most notably monosegmental morphemes in Hoocąk, which, like Totoli, has a relatively small inventory of mono-graphemic morphemes. A small inventory of morphemes could still be highly frequent, but in Hoocąk in particular, mono-graphemic morphemes are quite infrequent. There are only 185 tokens in the corpus, the vast majority appearing to be epenthetic vowels occurring between consonant-final verb roots and consonant-initial suffixes (93% of tokens).

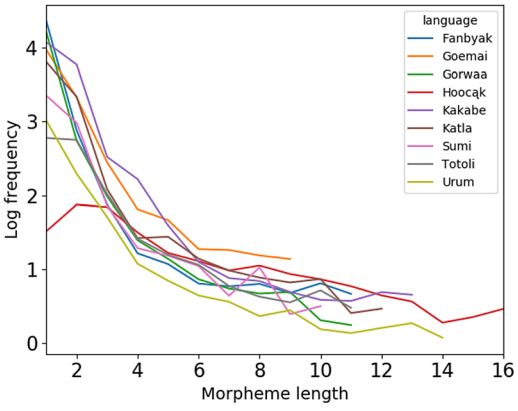

Figure 2: Log frequency of morpheme lengths by language (excluding morpheme lengths with fewer than 10 carrier word types for the purpose of visualization).

Figures 3 and 4 visualize the effect of Menzerath's Law in the nine corpora, excluding numbers of morphemes for which there were not at least 10 attested word types. Figure 3 shows the relationship between the number of morphemes in a word and the mean length of those morphemes. For each language, regardless of its morphological complexity, there is a clear pattern of longer morphemes for words containing fewer morphemes, with the steepest decline between 1 and 2 morphemes for almost all languages, and a more gradual decline thereafter. This decrease is very nearly monotonic across all languages.

Note that much of the effect shown in Figure 3 might be reducible to Zipf's law in the following sense: one should expect typically only one, relatively long and non-frequent lexical root per word, plus an increasing number relatively short and frequent affixes. However, the multivariate analysis presented below shows that Menzerath's law has an independent effect on word length.

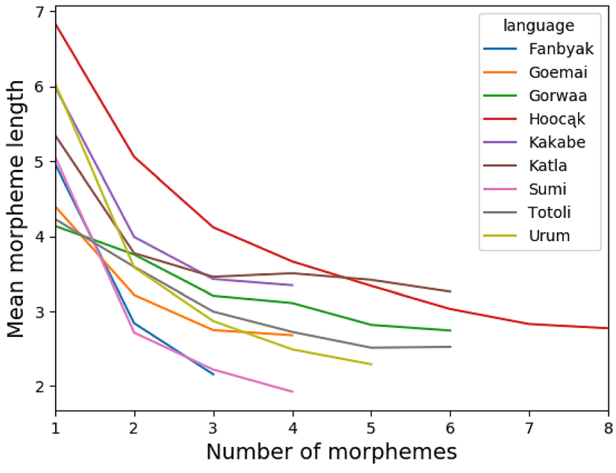

Figure 3: Mean morpheme length by number of morphemes in a carrier word (excluding morpheme lengths with fewer than 10 word types for the purpose of visualization).

Figure 4 illustrates the second measure of Menzerath's Law, plotting the mean length of morphemes in a word against the graphemic length of the carrier word (minus the morpheme itself). We see the same general pattern: longer words have shorter mean morpheme lengths. Compared to the number of morphemes in Figure 3, however, this pattern is not as robust. More isolating languages, like Fanbyak, Goemai, and Sumi, show only minimal decrease in mean morpheme length as word length increases, while more synthetic languages, like Hoocąk, Urum, and Katla, exhibit a clearer Menzerathian effect.

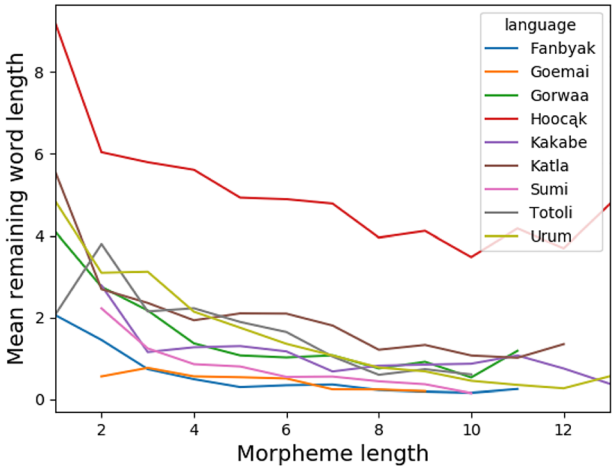

Figure 4: Mean morpheme length by remaining word length in a carrier word (excluding morpheme lengths with fewer than 10 word types for the purpose of visualization).

Because Menzerath's Law's effect is most clearly instantiated when there is a one-to-many relationship between sub-elements and the carrier element, the high number of mono-morphemic units presents a problem. To account for this, an ordinary least squares linear regression was run on all morphemes in the nine corpora that occurred in at least one multi-morphemic word (referred to hereafter as Multi-morph), using morpheme length as a dependent variable and the three above-mentioned as independent variables. Additionally, to account for the effect of mono-morphemic words as well, a separate regression was run using all morphemes in the nine corpora (referred to as All-morph). As expected, there was substantial correlation between the two measures of Menzerath's Law: 0.625 for the Multi-morph regression and increasing to 0.77 for the All-morph regression, due to the inclusion of mono-morphemes. Variance inflation factor was also calculated for independent variables in each regression, and all scores were below 2 for the Multi-morph regression, and below 2.5 for the All-morph regression, indicating that collinearity was within an acceptable range. Removing mono-morphemic morphemes, predictably, reduced a greater proportion of morphemes for isolating languages than for synthetic languages, as seen in Table 2.

Table 2: Number of morpheme types used in Multi-morph and All-morph regressions.

Language	All-morph	Multi-morph
Fanbyak	998	290
Goemai	1,146	493
Kakabe	2,031	1,118
Sumi	1,260	580
Totoli	1,273	828
Katla	1,467	1,071
Gorwaa	1,562	828
Urum	2,756	1,378
Hoocąk	2,093	1,361

In a first analysis, we ran regression models using graphemic morpheme length as a dependent variable, graphemic length of carrier word and number of morphemes in carrier word as Menzerathian independent variables, and morpheme frequency normalized by corpus length as a Zipfian independent variable. These achieve R^2 values of 0.126 and 0.098 respectively for the Multi-morph and All-morph models. Following inspection of residuals, an alternative model using log-transformed morpheme frequency was adopted, which improved residual distribution.

In the Multi-morph regression, we tested the influence of Zipf's and Menzerath's Laws on the lengths of morphemes, looking only at morphemes that participate in multi-morphemic carrier words. We found that the independent variables predicted 17.5% of the variance in morpheme length: $F = 221.2$ ($p < 0.001$); $R^2 = 0.175$; Morph_length = 5.875 – 0.262 × word_len – 0.400 × num_morph – 0.541 × norm_freq – 0.101 × word_len:norm_freq. The factors word_len, num_morph, and norm_freq were significant at $p < 0.001$, which indicates that all Zipfian and Menzerathian predictors contributed significantly to morpheme length. There was also a significant negative interaction between word_len and norm_freq, at $p < 0.05$, which means that length of the carrier word and morpheme frequency each decreased the effect of the other, when present.

In the All-morph regression, we performed the same test, but looking at all morphemes in the corpora, finding a reduced amount of variance explained, at 10.3%: $F = 223.2$ ($p < 0.001$); $R^2 = 0.103$; Morph_length = 5.701 – 0.267 × word_len – 0.551 × norm_freq – 0.151 × word_len:num_morph – 0.081 × num_morph:-norm_freq + error. Factors word_len and norm_freq were significant at $p < 0.001$, indicating that the Zipfian predictor contributed significantly to the model, but among the Menzerathian predictors, only word length was significant on its own. There were also significant negative interactions for word_len:num_morph ($p < 0.001$) and num_morph:norm_freq ($p < 0.05$), which means that an increased number of morphemes had a negative effect on morpheme length in the presence of either of the other two factors.

Interaction plots are shown in Figure 5 to illustrate the nature of the interactions in each model. Each interaction is between two continuous factors, so one variable for each plot has been converted to a binary factor of high and low (using quantile-based discretization, meaning the high and low groups of that factor are of equal size); the *y*-axis is the dependent variable (morpheme length) and the *x*-axis shows a normalized range of that independent variable. For the All-morph model, the number of morphemes influences how both morpheme frequency and mean word length relate to morpheme length. In both cases, carrier words with a high number of morphemes show a stronger negative relationship than those with a low number of morphemes. For the Multi-morph model, it is the graphemic length of the carrier word that affects the relationship between morpheme frequency and morpheme length, but again it is longer words that show a stronger negative relationship.

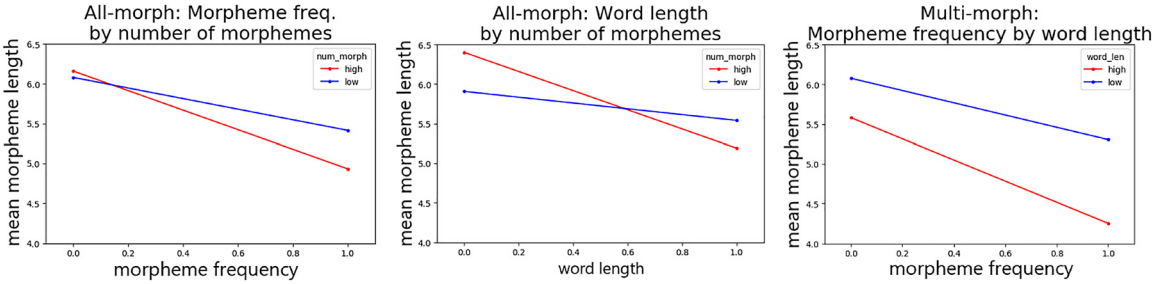

Figure 5: Interaction plots for All-morph and Multi-morph regressions.

To explore the variation in how the three factors correspond to morpheme length within each language, correlations were run for the dependent factor on each independent factor for each language. These results for multi-morphemic morphemes are summarized in Table 3, below, with significance scores calculated via Pearson's correlation coefficient. Partial regression plots are shown in Figure 6, where the strength of the effect of morpheme frequency can be seen in both regressions.

Table 3: Correlations between morpheme length and independent variables for Multi-morph and All-morph regressions, ordered by synthesis index (see Table 1).

	Multi-morph			All-morph		
Language	**word_len**	**num_morph**	**norm_freq**	**word_len**	**num_morph**	**norm_freq**
Fanbyak	−0.22***	−0.19**	−0.38***	−0.28***	−0.26***	−0.28***
Goemai	−0.22***	−0.13**	−0.30***	−0.19***	−0.17***	−0.23***
Kakabe	−0.07*	−0.06 (n.s.)	−0.29***	−0.11***	−0.05*	−0.25***
Sumi	−0.25***	−0.36***	−0.29***	−0.28***	−0.32***	−0.24***
Totoli	−0.28***	−0.16***	−0.26***	−0.29***	−0.22***	−0.26***
Katla	−0.15***	−0.15***	−0.26***	−0.16***	−0.15***	−0.24***
Gorwaa	−0.34***	−0.09**	−0.35***	−0.28***	−0.10***	−0.30***
Urum	−0.33***	−0.25***	−0.26***	−0.36***	−0.34***	−0.26***
Hoocąk	−0.26***	−0.24***	−0.20***	−0.11***	−0.09***	−0.14***
Mean	−0.24	−0.18	−0.29	−0.23	−0.19	−0.24

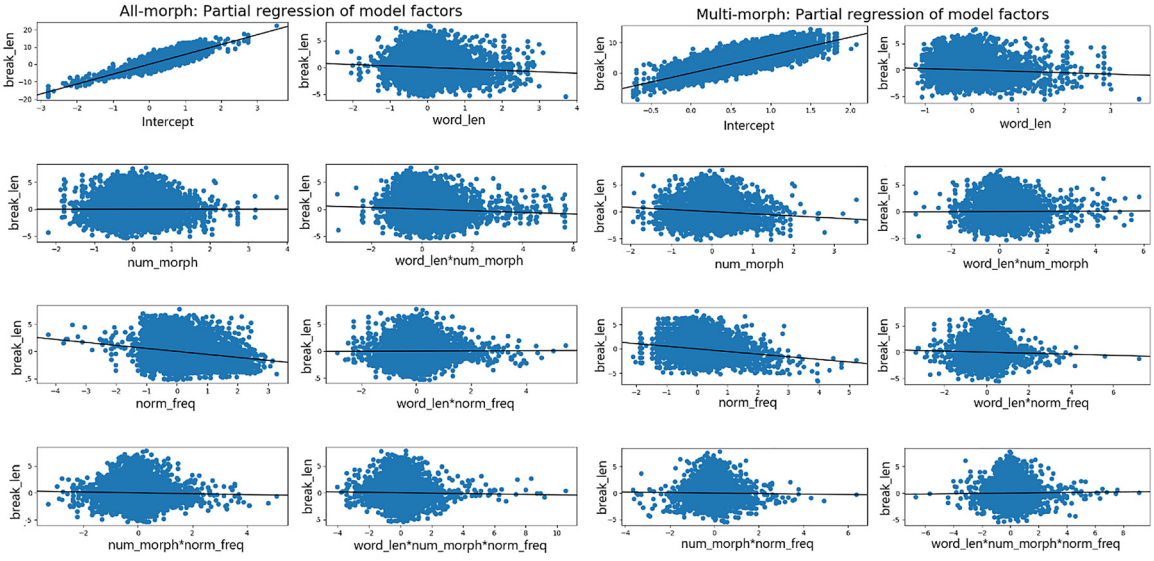

Figure 6: Partial regressions of model factors for Multi-morph and All-morph regressions.

4 Discussion

We saw in Figure 2 that Zipf's Law of Abbreviation operates at the morpheme level in all nine of the languages in our corpora, with more frequent morphemes being shorter than less frequent morphemes. This adds results from a typologically diverse sample of languages to the body of evidence showing that Zipf's Law is a basic principle of human language. These languages range from mostly isolating (Fanbyak) to fairly synthetic (Hoocąk), but all follow the same general pattern. Figures 3 and 4 illustrate that Menzerath's Law is also operating on morpheme lengths in all nine of the languages in the sample. As the number of morphemes in the carrier word increases, the mean length of those morphemes decreases, and as the mean remaining word length increases, the morpheme length also decreases.

In the Multi-morph regression, which predicted morpheme length among morphemes that participated in multi-morphemic words, we found that Zipf's and Menzerath's laws captured 17.5% of the variance in morpheme length. Both frequency for Zipf and carrier word length for Menzerath were significant predictors. Of the two laws, Zipf's Law was more predictive that Menzerath's Law. Removing mean number of morphemes per carrier word reduced the explained variance to 16.0%, while removing mean length of carrier word reduced explained variance to 16.7%. For comparison, removing morpheme frequency reduced the explained variance to 7.3%. Between the two measures of Menzerath's Law, the mean number of morphemes in the carrier word thus contributed only slightly more to the variance accounted for by the model. We also found a significant, though small, negative interaction between morpheme frequency and mean word length: there is a negative relationship between morpheme frequency and morpheme length, but this is more pronounced for longer words than for shorter words.

In the All-morph regression, which predicted morpheme length among all morphemes in the corpora, we found that the two laws capture only around 10.3% of the variance in morpheme length. Zipf's Law remains the more powerful predictor of the two laws. For Menzerath's Law, only mean word length is a predictive factor, while mean number of morphemes is only apparent in interactions. Removing mean word length from the model reduces explained variance to 9.5%, and morpheme frequency to 4.7%. Overall, the reduced amount of explained variance seems to be related to the reduced predictive power of mean number of morphemes, given the fact that many more morphemes in this sample have a value of 1 for this variable. There are also significant interactions between mean number of morphemes and both morpheme frequency and mean word length, both negative. In both cases, the negative relationship between morpheme length and morpheme frequency, or mean word length, is stronger for words with a higher number of morphemes than for words with a low number. This is likely driven by the prevalence of mono-morphemic words included in the All-morph regression.

In the Multi-morph model, Zipf's and Menzerath's Laws explain 17.5% of the variance of morpheme length. This may seem like a small amount, but if we consider the complexity of the linguistic system and the wealth of pressures that affect morphemic structure, it is not insubstantial. These laws are very general constraints, that are applied on top of (and often in spite of) language-specific grammatical, phonological, and prosodic constraints. It may also be that even stronger effects than those found for frequency here could be obtained if contextual predictability was measured. As mentioned above, recent research showed that Zipfian effects on word length are more closely correlated with predictability in context rather than frequency, although both measures are strongly correlated (Piantadosi et al. 2011, or see Gibson et al. 2019 for a broader survey). To what extent such explanations will work for morpheme length will have to be determined by future research, which will have to be based on corpora much larger than the ones available for the languages used in the current study, because reliable predictability measures requires more data than frequency measures.

To examine in more detail how Zipf's and Menzerath's Laws vary across languages, Table 3 showed the correlations between the independent variables and morpheme length. Overall, morpheme frequency shows the strongest correlation particularly for multi-morphemic words, while mean word length is a close competitor.

We see some indications of morphological type on the results. For morphemes occurring in multi-morphemic words, there is a slight decrease in the correlation between morpheme frequency and morpheme length as the morphological complexity increases, from −0.38 with an isolating language like Fanbyak, and −0.20 with a highly synthetic language like Hoocąk. However, most intermediate languages show much less variation, so this may be a characteristic only of languages at extremes of the synthesis spectrum. Gorwaa, with a correlation of −0.35, is an outlier, but it is also the only synthetic language that is predominantly fusional, which may also explain the disparity between its high correlation for word mean word length, and lower correlation for mean number of morphemes, relative to other synthetic languages. Relationships between morphological types and the measures of Menzerath's Law are less clear.

5 Conclusion

We examined the cross-linguistic distributions of morpheme length in nine languages, and cross-linguistic relationship between morpheme length and two universal laws predicted to explain this distribution. We looked at one measure of Zipf's Law of Abbreviation (morpheme frequency) and two measures of Menzerath's Law (mean grapheme length of words the morpheme appears in, minus the grapheme length of the morpheme itself, and mean number of morphemes in the morpheme's carrier words).

From a perspective of language efficiency, we saw that there were greater trade-offs for morpheme lengths in the Zipfian domain (frequency) than in the Menzerathian domain (length of the carrier unit), regardless of whether we included morphemes that do not participate in multi-morphemic words. This suggests that, for morphemes, global frequency plays a greater role in determining morpheme length than word-level complexity effects. However, Menzerathian effects, although they have received much less attention than Zipfian effects, are also clearly observed across all languages studied here. This calls for a reconsideration of the role of Zipfian efficiency in modelling language efficiency (e.g., Haspelmath 2018), as it is clearly only one relevant factor, Menzerathian being another.

The current study has also demonstrated both the usefulness and feasibility of a cross-linguistic approach to investigating morpheme length. Future work in this area can now expand to more languages, making use of newly available resources on a variety of languages.

References

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

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

Dryer, Matthew. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257–292.

Fedurek, Pawel, Klaus Zuberbühler & Stuart Semple. 2017. Trade-offs in the production of animal vocal sequences: Insights from the structure of wild chimpanzee pant hoots. *Frontiers in Zoology* 14(1). 50.

Fenk, Auguste & Gertraud Fenk-Oczlon. 1993. Menzerath's law and the constant flow of linguistic information. In Reinhard Köhler & Burghard B. Rieger (eds.), *Contributions to quantitative linguistics*, 11–31. Dordrecht: Springer.

Ferrer-i-Cancho, Ramon, Antoni Hernández-Fernández, David Lusseau, Govindasamy Agoramoorthy, Minna J. Hsu & Stuart Semple. 2013. Compression as a universal principle of animal behavior. *Cognitive Science* 37(8). 1565–1578.

Franjieh, Michael. 2018. Fanbyak corpus (Deposit IDs: 0131, 0387). London: ELAR.

Gerlach, Rainer. 1982. Zut Überprüfung des Menzerathschen Gesetzes im Bereich der Morphologie. *Glottometrika* 4. 95–102.

Gibson, Edward, Richard Futrell, Steven T. Piandadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen & Roger Levy. 2019. How efficiency shapes human language. *Trends in Cognitive Sciences* 23. 389–407.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2020. Glottolog 4.3. Jena: Max Planck Institute for the Science of Human History. https://doi.org/10.5281/zenodo.4061162 (accessed 05 March 2021).

Hartmann, Iren. 2004. Hoocąk text corpus. Nijmegen: The Language Archive. Available at: https://hdl.handle.net/1839/bf362b19-2cd6-4bee-af44-a46446077875.

Harvey, Andrew. 2016. *The Gorwaa noun phrase: Toward a description of the Gorwaa language* (Deposit ID: 0404). London: ELAR.

Haspelmath, Martin. 2018. Explaining grammatical coding asymmetries: Form-frequency correspondences and predictability. Manuscript. Leipzig, ms. Available at: https://ling.auf.net/lingbuzz/004531.
Haspelmath, Martin & Andrea Sims. 2013. *Understanding morphology*. New York, NY: Routledge.
Heesen, Raphaela, Catherine Hobaiter, Ramon Ferrer-i-Cancho & Stuart Semple. 2019. Linguistic laws in chimpanzee gestural communication. *Proceedings of the Royal Society B* 286(1896). 20182900.
Hellwig, Birgit. 2003. *Goemai texts* (Deposit ID: 0003). London: ELAR.
Hellwig, Birgit. 2007. A Documention of Tabaq, a Hill Nubian language of the Sudan. London: ELAR, in its sociolinguistic context (Deposit ID: 0200). Available at: https://elar.soas.ac.uk/Collection/MPI143018.
Kisler, Thomas, Uwe Reichel & Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45. 326–347.
Köhler, Reinhard. 1982. Das Menzerathsche Gesetz auf Satzebene. *Glottometrika* 4. 103–113.
Köhler, Reinhard. 1984. Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika* 6. 177–183.
Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, MA: The M.I.T. Press.
Leto, Claudia, Winarno S. Alamudi, Nikolaus P. Himmelmann, Jani Kuhnt-Saptodewo, Sonja Riesberg & Hasan Basri. 2010. DoBeS Totoli documentation. DoBeS Archive MPI Nijmegen. Available at: http://www.mpi.nl/DOBES/.
Mačutek, Ján, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann Law in syntactic dependency structure. *International conference on dependency linguistics (Depling)*, 4, 100–107. Linköping: University Electronic Press.
Mandelbrot, Benoit. 1953. An informational theory of the statistical structure of language. *Communication Theory* 84. 486–502.
Menzerath, Paul. 1928. Über einige phonetische Probleme. In *Actes du premier congrès international de linguistes*. Leiden: Sijthhoff.
Milička, Jiří. 2014. Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics* 21(2). 85–99.
Miller, George. A. 1957. Some effects of intermittent silence. *American Journal of Psychology* 70(2). 311–314.
Paschen, Ludger, François Delafontaine, Cristoph Draxler, Susanne Fuchs, Matthew Stave & Frank Seifart. 2020. Building a time-aligned cross-linguistic reference corpus from language documentation data (DoReCo). In *Proceedings of the 12th language resources and evaluation conference (LREC)*, 2657–2666. https://www.aclweb.org/anthology/2020.lrec-1.324.
Piantadosi, Steven. T., Harry Tily & Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences* 108(9). 3526–3529.
Pustet, Regina & Altmann Gabriel. 2005. Morpheme length distribution in Lakota. *Journal of Quantitative Linguistics* 12(1). 53–63.
Skopeteas, Stavros. & Violeta Moisidi. 2011. Texts: Urum narrative collection (Working papers of the Urum documentation project). Bielefeld: University of Bielefeld. Available at: http://urum.lili.uni-bielefeld.de/.
Smith, F. J. & K. Devine. 1985. Storing and retrieving word phrases. *Information Processing & Management* 21(3). 215–224.
Strunk, Jan., Seifart Frank, S. Danielsen, Hartmann Iren, Brigitte Pakendorf, Søren Wichmann, Alena Witzlack-Makarevich & Balthasar Bickel. 2019. *Determinants of phonetic word duration in ten language documentation corpora: Word frequency, complexity, position, and part of speech*. Submitted manuscript. University of Cologne, ms.
Teo, Amos. 2013. *Documenting traditional agricultural songs and stories of the Sumi Nagas*. (Deposit ID: 0128). London: ELAR.
Teupenhayn, R. & Altmann. Gabriel. 1984. Clause length and Menzerath's law. *Glottometrika* 6. 127–138.
Vydrina, Alexandra. 2013. *Description and documentation of the Kakabe language*. (Deposit ID: 0228). London: ELAR.
Zipf, George Kingsley. 1935. *The psycho-biology of language: An introduction to dynamic philology*. Houghton Mifflin. Reprinted 1968. Cambridge, MA: The M.I.T. Press.
Zipf, George K. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.