Germán Coloma

The Menzerath-Altmann Law in a Cross-Linguistic Context

Abstract

This paper attempts to contrast two alternative formulae for the Menzerath-Altmann law, using data from two linguistic measures (words per clause and phonemes per word) for the same text translated into 50 different languages. The alternative formulae are the traditional power function, and a recently proposed hyperbolic function. The estimations are modified to control for genetic and geographic factors, and for the presence of possible endogeneity between the related variables. None of these significantly alter the basic results, which show a slight preference for the power function over the hyperbolic one.[*]

1. Introduction

The Menzerath-Altmann law states that the length of a linguistic construct is an inverse function of the length of the construct's constituents. Originally proposed by Menzerath (1954), this law was reformulated by Altmann (1980) as a power function that can be written in the following way:

(1) $y = a \cdot x^b$;

[*] I thank Gabriel Altmann, Mariana Conte Grand, Stefan Gries, Reinhard Kohler, Fermín Moscoso, and two anonymous referees, for their useful comments. I especially thank Mirka Rauniomaa, the editor of the *SKY Journal of Linguistics*, and I also thank Federico Cápula, Valeria Dowding, Helen Eaton, John Esling, Sameer Kahn, Kevin Schäfer and Justin Watkins, for their help in finding some of the sources used in this article. Valeria Dowding also helped me with the English. Part of the research for this paper was conducted while I was working as a visiting scholar at the University of California, Santa Barbara.

where *y* is the average length of a linguistic construct, measured in its constituents, *x* is the average length of the construct's constituents, measured in their subconstituents, and *a* and *b* are parameters.[1]

Many applications of this law exist for linguistic data. Menzerath's and Altmann's were related to word length and syllable length, but other applications include relationships between sentence length and word length (Teupenhayn & Altmann 1984), between sentence length and clause length (Kulacka 2010), and between word length and number of distinct words (Eroglu 2013). Although most analyses have been performed in single-language contexts (i.e., using texts written in the same language), the Menzerath-Altmann law has also been used to explain phenomena that occur in cross-linguistic environments (see, for example, Fenk-Oczlon & Fenk 1999).[2]

In a recent paper (Milicka 2014), it is argued that the traditional (power function) formula for the Menzerath-Altmann law can be improved by using a hyperbolic alternative, written in the following way:

$$(2) \qquad y = a + \frac{b}{x} \quad .$$

This formula is supposed to fit some datasets better and to have a more intuitive explanation, related to a trade-off between plain information and structure information (Köhler 1984).[3]

When one applies the Menzerath-Altmann law in a single-language context, its results are linked to situations in which language users concentrate information in a small number of more complex units, as opposed to situations in which they prefer to use a larger number of simpler units in order to convey that information. If the unit is the word, then the more complex words are the ones that have more morphemes, or syllables, or phonemes, and the simpler ones are the ones that have fewer of those elements.

---

[1] In fact, Altmann's formula also includes an additional exponential term ($e^{c \cdot x}$). This term disappears when we solve the formula as a differential equation.

[2] Some applications of the Menzerath-Altmann law have even gone further, and tested for the existence of similar relationships in areas that are away from linguistics. See, for example, Boroda & Altmann (1991) for musical texts, and Ferrer & Forns (2010) for genomes.

[3] The traditional Menzerath-Altmann law power function, however, has also had several proposed theoretical explanations. Eroglu (2014), for example, has interpreted it as a particular case of a statistical mechanical organization.

When we generalize the use of the Menzerath-Altmann law to a cross-linguistic context, then the interpretation of its results can be related to the possible existence of complexity trade-offs between language sub-systems (Fenk-Oczlon & Fenk 2008). As language structures typically vary more when we consider different languages instead of different uses within the same language, those complexity trade-offs can be seen as representative of particular strategies that languages use to communicate the same information, which could imply using more units (e.g., more words per clause) or larger units (e.g., more phonemes per word). This can also be linked to conceptual characteristics of language structure, for example, syntax (which is expected to be more complex if a language uses more words per clause) and morphology (which is expected to be more complex if a language uses more phonemes per word).[4]

In the following sections of this paper we will proceed to compare the implications of formulae 1 and 2 for the Menzerath-Altmann law in a cross-linguistic context. In order to do that, we will use a sample of 50 languages for which we have data for the same text. In each case, we will calculate the number of phonemes per word and the number of words per clause, and see which version of the law fits the data better. Our comparison will be later improved by running two different specification tests, and by including the possible effect of two categorical variables: location and genetic affiliation of the languages. We will also include a correction related to the possible endogeneity of the phoneme/word ratio as an explanatory variable for the word/clause ratio.

2. Description of the data

The text from which we derive the results presented in this paper is the fable known as "The North Wind and the Sun", attributed to Aesop, which is a short story used by the International Phonetic Association (IPA) as a "specimen" or model to illustrate the phonetics of a considerable number of languages. This text has the advantage of being clearly described in terms of its constituting phonemes, words and clauses, and it is also immediately comparable across languages. For example, the (Standard Southern British) English version of "The North Wind and the Sun" is the following:

[4] For other alternatives concerning the empirical measurement of morphological and syntactic complexity, see Kettunen, McNamee and Baskaya (2010) and Szmrecsányi (2004).

> The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Its corresponding phonemic transcription is this:

> ðə nɔːθ wɪnd ən ðə sʌn wə dɪspjuːtɪŋ wɪtʃ wəz ðə strɒŋgə | wən ə trævlə keɪm əlɒŋ ræpt ɪn ə wɔːm kləʊk || ðəɪ əgriːd ðət ðə wʌn hu fɜːst səksiːdɪd ɪn meɪkɪŋ ðə trævlə teɪk hɪz kləʊk ɒf ʃʊd bi kənsɪdəd strɒŋgə ðən ði ʌðə || ðən ðə nɔːθ wɪnd bluː əz hɑːd əz i kʊd | bət ðə mɔː hi bluː ðə mɔː kləʊsli dɪd ðə trævlə fəʊld hɪz kləʊk əraʊnd hɪm | ænd ət lɑːst ðə nɔːθ wɪnd geɪv ʌp ði ətempt || ðən ðə sʌn ʃɒn aʊt wɔːmli | ænd əmiːdiətli ðə trævlə tʊk ɒf ɪz kləʊk || n səʊ ðə nɔːθ wɪn wəz əblaɪdʒd tu kənfes ðət ðə sʌn wəz ðə strɒŋgr əv ðə tuː ||

If we count the number of clauses, words and phonemes in this text, we can find that it consists of 9 clauses,[5] 113 words and 383 phonemes. It therefore has 3.39 phonemes per word and 12.56 words per clause. The same calculations can be made for other available cases with different relationship between those ratios. For example, the Turkish version of "The North Wind and the Sun" has a larger number of phonemes per word (equal to 6.53) than the English version, but its number of words per clause (equal to 7.33) is much smaller.

In order to perform our analysis, we selected a sample of 50 languages for which we found versions of the abovementioned text in either the *Handbook of the International Phonetic Association* (IPA 1999) or in the series of "Illustrations of the IPA", published by the *Journal of the International Phonetic Association*. That sample consists of ten languages from each of the five areas in which we divided the world, which are America (Sahaptin, Apache, Chickasaw, Seri, Trique, Zapotec, Quichua, Shiwilu, Yine and Mapudungun), Europe (Portuguese, Spanish, Basque, French, Irish, English, German, Russian, Hungarian and Greek), Africa (Tashlhiyt, Nara, Dinka, Amharic, Sandawe, Bemba, Hausa, Igbo, Kabiye and Temne), West Asia (Georgian, Turkish, Hebrew, Arabic, Persian,

[5] The concept of clause that we use for this calculation is based on the number of pauses marked in the phonemic text, and not on syntactic considerations. This allows making comparisons easier when we deal with different languages.

Tajik, Nepali, Hindi, Bengali and Tamil) and East Asia (Japanese, Korean, Mandarin, Cantonese, Burmese, Thai, Vietnamese, Malay, Tausug and Arrernte). In this division, Australia (where the Arrernte language is spoken) is considered as a part of East Asia.

Some language families are represented by more than one language. For example, our sample includes 13 Indo-European languages (Portuguese, Spanish, French, Irish, English, German, Russian, Greek, Persian, Tajik, Nepali, Hindi and Bengali), 5 Afro-Asiatic languages (Tashlhiyt, Amharic, Hausa, Hebrew and Arabic), 4 Niger-Congo languages (Bemba, Igbo, Kabiye and Temne) and 3 Sino-Tibetan languages (Mandarin, Cantonese and Burmese).

The complete dataset used is reproduced in Appendix 1. In it we can see that the average number of phonemes per word for the whole sample is equal to 4.76, with a minimum value of 2.85 (that corresponds to the Vietnamese language) and a maximum value of 8.87 (that corresponds to Yine, which is an Arawakan language spoken in Peru). The maximum value for the word/clause ratio, conversely, is reported for the Irish language (and is equal to 18.43), while the minimum value for that ratio is 5.70 (and corresponds to Chickasaw, a Muskogean language spoken in the United States), in a context where the average number of words per clause is equal to 10.19.

In order to calculate the numbers mentioned in the previous paragraphs, we first had to define the number of words and phonemes in each version of "The North Wind and the Sun". To do that, we basically followed the criteria used by the authors that wrote the corresponding illustrations of the IPA, who in all cases use a certain standard to separate the text into words, and the words into phonemes. We also applied some unifying criteria, though. For example, long, short, oral and nasal vowels were considered as different phonemes when length or nasalization were distinctive in a certain language, but diphthongs were always considered as a combination of two phonemes within the same syllable. Affricate consonants and other "double articulations" were also considered as separate phonemes when appropriate, while "geminate consonants" were always considered to be a combination of two (identical) consecutive phonemes.

Phonemes per word and words per clause have a relatively large negative correlation between them in this sample. Measured by the (Pearson) product-moment coefficient, that correlation is equal to -0.7182. Taking into account the fact that it is obtained from a sample of 50

observations (with 48 degrees of freedom), its corresponding t-statistic is equal to -7.1515. This statistic generates a p-value equal to 0.000000004 (which makes it statistically different from zero at any reasonable probability level).

3. Application of the alternative formulae for the Menzerath-Altmann law

In order to test the relative performance of Equations (1) and (2) for the Menzerath-Altmann law, we will run regressions using the data described in section 2. Those regressions will be based on the following functions:

(3) $Ln(Word/Clause) = c(1) + c(2) * Ln(Phon/Word)$;

(4) $Word/Clause = c(1) + c(2) * [1/(Phon/Word)]$;

which are linear versions of the original equations for the case where the independent variable is a logarithmic or an inverse transformation of the phoneme/word ratio (*Phon/Word*), and the dependent variable is the word/clause ratio (*Word/Clause*) or a logarithmic transformation of it.

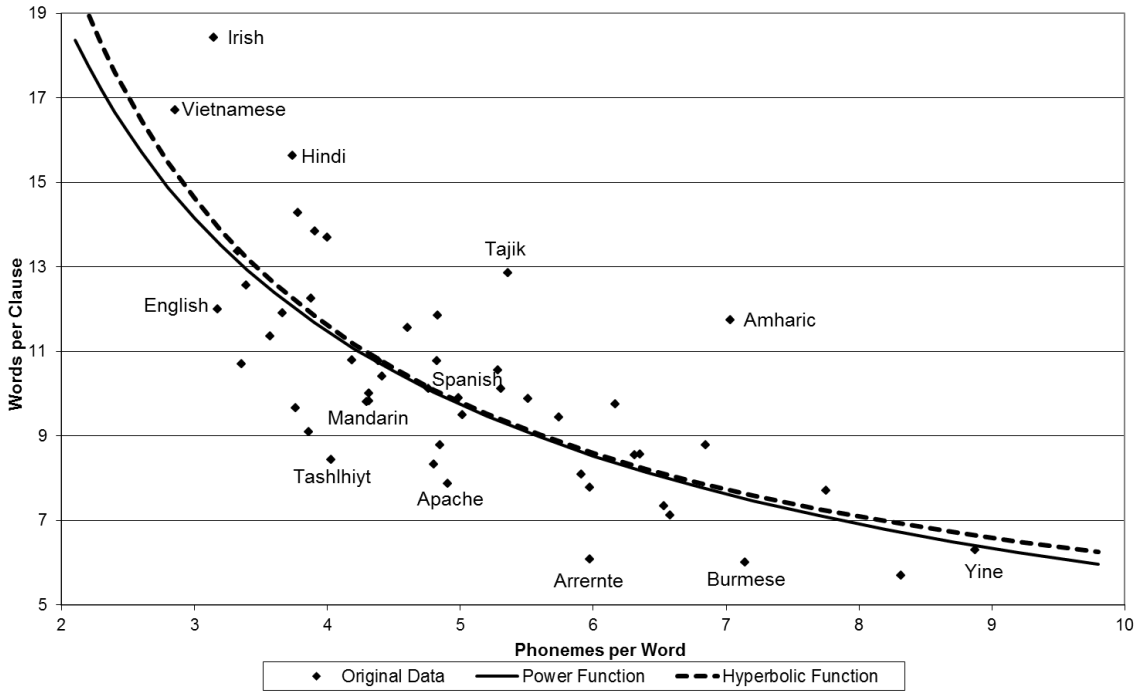
Table 1. Regression results from OLS estimation

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Power function				
Constant (*c*(1))	3.4528	0.1392	24.7963	0.0000
Phon/Word (*c*(2))	-0.7310	0.0874	-8.3637	0.0000
R-squared	0.5931			
Adjusted R^2	0.5846			
Hyperbolic function				
Constant (*c*(1))	2.5735	0.9878	2.6052	0.0122
Phon/Word (*c*(2))	36.1152	4.4371	8.1394	0.0000
R-squared	0.5799			
Adjusted R^2	0.5711			

The main results for those regressions, run using ordinary least squares (OLS), are presented in Table 1. In it we can see that both functions generate a relatively good fit for the data. The estimated regression coefficients are also highly significant, and they have the expected signs. They both imply a negative relationship between *Word/Clause* and *Phon/Word*. Based on their R^2 coefficients, one can also find that the fit of the power function (R^2 = 0.5931) is slightly better than the one obtained

under the hyperbolic function ($R^2 = 0.5799$). Both specifications also have a better fit than the one that could be obtained under a simpler linear regression. That specification would have produced an R^2 coefficient equal to 0.5158.

Figure 1. Power and hyperbolic regression curves



The power-function and hyperbolic-function regression equations can also be graphed in a diagram in which one represents the different language observations in terms of phonemes per word versus words per clause. That is what appears in Figure 1, where we see that the hyperbolic equation predicts a higher value of *Word/Clause* for any possible value of *Phon/Word*. This generates a better fit for 24 languages (e.g., Irish, Vietnamese, Hindi, Tajik, Amharic) but a worse fit for the remaining 26 languages (e.g., English, Mandarin, Apache, Burmese, Arrernte).

4. Specification tests

To have a more precise way to deal with the relative advantages of the power function and the hyperbolic function as alternative specifications of the Menzerath-Altmann law, we can run some tests aimed at comparing if the results from one regression explain phenomena that remain unexplained by the other regression. One of such tests has been proposed by Davidson

and MacKinnon (1981), and is also known as the J test. This consists of running regressions like the following:

(5) $Ln(Word/Clause) = c(1) + c(2) * Ln(Phon/Word) + c(3) * Ln(WC2fitted)$;

(6) $Word/Clause = c(1) + c(2) * [1/(Phon/Word)] + c(3) * WC1fitted$;

where *WC1fitted* and *WC2fitted* are the values estimated for the word/clause ratio by the regressions run under Equations (3) and (4). The idea of this test is to analyze if the behavior of the dependent variable that is explained under one model can help to improve the estimation under the alternative model, and the basic element to evaluate this is the statistical significance of the coefficients labeled as *c*(3) in Equations (5) and (6).

Table 2. Results from J-test regressions

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Power function				
Constant (*c*(1))	1.2015	10.5024	0.1144	0.9094
Phon/Word (*c*(2))	-0.2540	2.2265	-0.1141	0.9096
WC fitted (*c*(3))	0.6482	3.0233	0.2144	0.8312
R-squared	0.5935			
Adjusted R^2	0.5762			
Hyperbolic function				
Constant (*c*(1))	7.0085	8.8345	0.7933	0.4316
Phon/Word (*c*(2))	93.5441	113.7574	0.8223	0.4151
WC fitted (*c*(3))	-1.6452	3.2564	-0.5052	0.6158
R-squared	0.5821			
Adjusted R^2	0.5644			

In Table 2, we can see the results for the regressions run under Equations (5) and (6). In both cases, the estimated value for *c*(3) is not significantly different from zero at any reasonable probability level ("p = 0.8312" and "p = 0.6158"). This indicates that the results generated by the power function cannot be appreciably improved by the factors taken into account by the hyperbolic function, while the opposite is also true (the factors taken into account by the power function cannot help to improve the estimation run under the hyperbolic function). Moreover, if we compare the adjusted R^2 coefficients that appear in Table 2 with the ones reported in Table 1, we can see that in both cases those coefficients have dropped, and this is another indication that the additional regressors do not help to improve the original results.

The J tests run on our two formulae for the Menzerath-Altmann law are examples of "non-nested tests", which consider the possible alternatives as competing models to be contrasted. In this case, we can also think of a "nested test", based on a general model that includes the power function and the hyperbolic function as special cases. The simplest of those models is the following:

(7) $y = a + b \cdot x^{c}$;

which in our case can be written as

(8) $Word/Clause = c(1) + c(2) * (Phon/Word)^{c(3)}$.

In order to estimate the parameters of a model like this, we need to run a non-linear regression like the one whose results appear in Table 3. The power function is therefore a particular case of Equation (8) for which it holds that "$c(1) = 0$", while the hyperbolic function is another particular case for which it holds that "$c(3) = -1$". The first of those restrictions in the parameter values can be tested by looking at the p-value of the corresponding coefficient (p = 0.1125) and it cannot be rejected at a 10% probability level. To test the restriction that "$c(3) = -1$" we have to run an additional test, the so-called "Wald test". That test produces a chi-square statistic (χ^2) for which it holds that "p = 0.4911", and this cannot be rejected at a 10% probability level, either.

Table 3. Results from a general OLS regression

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Constant ($c(1)$)	5.4108	3.3456	1.6173	0.1125
Multiplicative parameter ($c(2)$)	56.1031	42.9824	1.3053	0.1982
Power parameter ($c(3)$)	-1.6035	0.8765	-1.8294	0.0737
R-squared	0.5832			
Adjusted R^2	0.5655			

The results reported in Table 3 also show an adjusted R^2 coefficient which is equal to 0.5655. That coefficient is smaller than the ones reported in Table 1, and this is another indication that both the power function and the hyperbolic function can be used to explain the data, and that a general model that includes both of them is not efficient to improve the explanatory power of each of the most simple formulae.

5. Geographic and genetic factors

A possible explanation for some variation in the word/clause ratios that remains unexplained by Equations (3), (4) and (8) is the existence of some areal and genetic factors that may determine that the functional relationship between *Word/Clause* and *Phon/Word* is not the same for every language. In order to take into account some of those factors, we included two additional categorical variables, related to the five regions in which we divided the sample and to the four main language families represented. This is equivalent to introducing binary variables for four out of the five regions (Africa, America, West Asia and East Asia) and for the four main language families (Indo-European, Afro-Asiatic, Niger-Congo and Sino-Tibetan).

Table 4 shows the results of these new regressions, run under alternative power-function and hyperbolic-function specifications. Although the included binary variables are in general not significant individually, they are indeed helpful in improving the fit of the estimations, whose adjusted R^2 coefficients go up from 0.5846 to 0.6271 (for the power function) and from 0.5711 to 0.6053 (for the hyperbolic function). This improvement, however, has no effect in the ranking of the R^2 coefficients, which still shows the power function ahead of the hyperbolic function.

Table 4. Results from regressions with geographic and genetic factors

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Power function				
Constant (*c*(1))	3.2725	0.1721	19.011	0.0000
Africa (*c*(2))	0.0652	0.1064	0.6132	0.5432
America (*c*(3))	-0.0518	0.0968	-0.5349	0.5957
West Asia (*c*(4))	0.0270	0.0787	0.3437	0.7329
East Asia (*c*(5))	0.0354	0.1011	0.3497	0.7284
Indo-European (*c*(6))	0.1319	0.0838	1.5738	0.1234
Afro-Asiatic (*c*(7))	0.0480	0.0932	0.5151	0.6093
Niger-Congo (*c*(8))	-0.0082	0.1121	-0.0728	0.9423
Sino-Tibetan (*c*(9))	-0.1906	0.1087	-1.7533	0.0872
Phon/Word (*c*(10))	-0.6430	0.0972	-6.6143	0.0000
R-squared	0.6956			
Adjusted R^2	0.6271			

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Hyperbolic function				
Constant (*c*(1))	2.8134	1.3860	2.0298	0.0491
Africa (*c*(2))	0.7845	1.1411	0.6875	0.4957
America (*c*(3))	-0.4283	1.0350	-0.4138	0.6812
West Asia (*c*(4))	0.4113	0.8478	0.4851	0.6302
East Asia (*c*(5))	0.5002	1.0854	0.4608	0.6474
Indo-European (*c*(6))	1.3996	0.9001	1.5549	0.1279
Afro-Asiatic (*c*(7))	0.4090	1.0003	0.4089	0.6848
Niger-Congo (*c*(8))	-0.3426	1.2044	-0.2845	0.7775
Sino-Tibetan (*c*(9))	-1.8810	1.1667	-1.6122	0.1148
Phon/Word (*c*(10))	32.5966	4.9553	6.5781	0.0000
R-squared	0.6778			
Adjusted R^2	0.6053			

The newly estimated equations can also be subject to J tests, to see if the results from one regression explain some phenomena that remain unexplained by the other regression. In this case, the estimated additional coefficients have probability values that are equal to "p = 0.1340" for the coefficient that measures the effect of hyperbolic factors on the power-function equation, and to "p = 0.2721" for the one that measures the effect of power-function factors on the hyperbolic equation. These coefficients fail to be statistically significant at a 10% probability level.

The inclusion of geographical and genetic factors on the relationship between phonemes per word and words per clause, which appears in our alternative formulae for the Menzerath-Altmann law, also seems to reduce the absolute magnitude of that relationship. Comparing the coefficients that appear in Table 4 with the ones reported in Table 1, we see that the negative coefficient for the phoneme/word ratio in the power function drops from 0.73 to 0.64, which implies a 12% reduction. In the same fashion, the inclusion of geographic and genetic factors implies a reduction in the equivalent coefficient from the hyperbolic function from 36.12 to 32.60 (i.e., a 9.7% decrease). Nevertheless, the new coefficients are still very significant, since their probability values are both indistinguishable from zero.

6. Instrumental variables

When one performs a regression between two variables, it is implicitly assumed that the variable on the right-hand side of the equation (i.e., the independent variable) is the one that explains the behavior of the variable included on the left-hand side of the equation (i.e., the dependent variable), and not the other way round. This is a noticeable difference between regression and correlation analyses, since correlation is a symmetrical concept that assumes no particular causal direction from one variable to the other.

In the case under study in this paper, the logic of the Menzerath-Altmann law indicates that the nature of the constituents of a language (i.e., the number of phonemes per word) determines the structure of the higher-level construct (i.e., the number of words per clause). However, this causality is not completely clear, especially if we examine a cross-linguistic context where we can interpret the relationship between the two variables as a signal of the existence of a complexity trade-off. In that context, both the word/clause ratio and the phoneme/word ratio may be variables that are simultaneously determined by an external process.

To deal with this kind of endogeneity issues we can use instrumental variables, i.e., variables that are supposed to be related with the independent variable under analysis but have the property that they are determined exogenously (i.e., outside the statistical problem that we are analyzing). For this particular case, we have chosen six numerical variables, which come from the different languages' grammars (and not from the texts that we use to compute the number of words per clause and the number of phonemes per word). Those variables are the number of consonant phonemes in each language's inventory (Consonants), the number of vowel phonemes in that inventory (Vowels), and the number of distinctive tones that each language possesses (Tones),[6] together with the number of distinctive genders that nouns may have (Genders), the number of distinctive cases for those nouns (Cases), and the number of inflectional categories of the verbs (Inflections).[7] The values for the first three instrumental variables are taken from the same sources used to obtain the different versions of "The North Wind and the Sun" (i.e., from the

[6] This number is equal to one for languages in which tone is not distinctive, and equal to the number of distinctive tones for the case of tonal languages.

[7] The complete set of instrumental variables is reproduced in Appendix 2.

corresponding illustrations of the IPA). To impute values for the last three variables, conversely, we used the online version of the *World Atlas of Language Structures* (WALS), edited by Dryer and Haspelmath (2013).

In a case like this, one can use a procedure to include the instrumental variables in the estimation of the equation coefficients that is known as "two-stage least squares" (2SLS). It consists of a first stage in which the endogenous independent variable (in our case, *Phon/Word*) is regressed against all the instrumental variables, using ordinary least squares. Then there is a second stage in which the fitted values of that regression are included in the estimation of the actual equation that one wishes to regress (in our case, in each of the Menzerath-Altmann law equations), instead of the original values for the endogenous independent variable.[8]

To perform the first stage of this 2SLS procedure, we used a logarithmic specification in which the natural logarithm of *Phon/Word* was regressed against a constant, the four geographic binary variables, the four genetic binary variables, and the natural logarithms of the six additional instrumental variables. Then the fitted values from that regression were used to replace the original values for *Phon/Word*, in new regressions that followed the specifications stated in Equations (3), (4) and (8).

The results of these 2SLS regressions for the most basic models are reported in Table 5. The table shows that the corresponding R^2 coefficients are in all cases smaller than the ones reported in Tables 1 and 3. This has to do with the fact that an estimation that uses instrumental variables is always less efficient than another estimation that uses the original variables, although it can be more consistent (i.e., closer to the true values of the parameters that would be obtained if one knew the whole set of data that is generating the process under estimation). In this case, the results are in line with the estimations performed in the previous sections, in the sense that the coefficients are significantly different from zero and imply a negative relationship between phonemes per word and words per clause. Once again, the power function has an advantage in terms of goodness of fit over the hyperbolic function, both in the comparison between standard R^2 coefficients and between adjusted R^2 coefficients.

[8] This procedure was originally proposed by Basmann (1957). For a more complete explanation, see Davidson and MacKinnon (2003: ch 8).

Table 5. Regression results from 2SLS estimation

Concept	**Coefficient**	**Std Error**	**t-Statistic**	**Probability**
Power function				
Constant (*c*(1))	3.5860	0.1777	20.1763	0.0000
Phon/Word (*c*(2))	-0.8158	0.1120	-7.2828	0.0000
R-squared	0.5249			
Adjusted R^2	0.5150			
Hyperbolic function				
Constant (*c*(1))	2.1803	1.1879	1.8355	0.0726
Phon/Word (*c*(2))	38.2906	5.4203	7.0643	0.0000
R-squared	0.5097			
Adjusted R^2	0.4995			
General regression				
Constant (*c*(1))	-10.1589	53.9499	-0.1883	0.8514
Multiplic parameter (*c*(2))	38.6476	38.6782	0.9992	0.3228
Power parameter (*c*(3))	-0.4061	1.0557	-0.3847	0.7022
R-squared	0.5124			
Adjusted R^2	0.4917			

When we run a 2SLS regression for a general specification of the model, we obtain results that are considerably different from the ones that appear in Table 3 (i.e., from the results of the same regression under ordinary least squares). However, if we test for the reasonableness of the competing models, we obtain the same conclusions, which imply that the restrictions associated with the power-function model (*c*(1) = 0) and the hyperbolic model (*c*(3) = -1) are both insignificantly different from zero ("p = 0.8514" and "p = 0.5737").

The same 2SLS estimation procedure can be used for the more complex versions of the model that also include the geographic and genetic binary variables as regressors in the Menzerath-Altmann law equations. When doing so, we end up with new estimations for the phoneme/word ratio coefficients, which are not significantly different from the original ones reported in Table 4. The ranking of R^2 coefficients does not change, either, in the sense that the power-function version of the law has a better fit ($R^2 = 0.6156$) than the hyperbolic version ($R^2 = 0.5955$).

7. Concluding remarks

The main conclusion that can be drawn from the different analyses performed in this study is that there is no evidence that the hyperbolic specification of the Menzerath-Altmann law (Milicka 2014) fits the data better than the original power function (Altmann 1980). This conclusion is based on a dataset that uses the same text translated into 50 different languages, and it holds for a version of the Menzerath-Altmann law that postulates a relationship between phonemes per word and words per clause.

Both the power function and the hyperbolic function perform well to explain the strong negative correlation that exists between phonemes per word and words per clause in this context, since the coefficients obtained for the phoneme/word ratio as an explanatory variable of the word/clause ratio have the expected sign, and they are also significantly different from zero at a 1% probability level.

All the tests that were performed in order to evaluate the relative merits of the power function and the hyperbolic function show that the variation in the word/clause ratio that remains unexplained using one function is not further explained by the alternative function. They also show that, when we nest both models into a single more general function, the additional parameter becomes statistically insignificant.

The traditional power-function formula, however, always displays larger coefficients of determination than the newly proposed hyperbolic formula. This advantage in fitting the data appears when we use the most basic formulation of the model (i.e., when we run a simple regression for the word/clause ratio as a function of the phoneme/word ratio), when we include geographic and genetic factors, and also when we use instrumental variables (consonant inventory, vowel inventory, number of tones, number of cases, number of genders, and number of possible verbal inflections).

References

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

Basmann, Robert (1957) A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica* 25: 77–83.

Boroda, Moisei & Altmann, Gabriel (1991) Menzerath's law in musical texts. *Musikometrica* 3: 1–13.

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

Davidson, Russell & MacKinnon, James (1981) Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49: 781–793.
—— (2003) *Econometric Theory and Methods*. New York: Oxford University Press.
Dryer, Matthew & Haspelmath, Martin (2013) *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Eroglu, Sertac (2013) Menzerath-Altmann law for distinct word distribution analysis in a large text. *Physica A* 392: 2775–2780.
—— (2014) Menzerath-Altmann law: Statistical mechanical interpretation as applied to a linguistic organization. *Journal of Statistical Physics* 157: 392–405.
Fenk-Oczlon, Gertraud & Fenk, August (1999) Cognition, quantitative linguistics and systemic typology. *Linguistic Typology* 3: 151–177.
—— (2008) Complexity trade-offs between the subsystems of language. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language Complexity: Typology, Contact and Change*, pp. 43–65. Amsterdam: John Benjamins.
Ferrer, Ramón & Forns, Nuria (2010) The self-organization of genomes. *Complexity* 15: 34–36.
IPA (1999) *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.
Kettunen, Kimmo; McNamee, Paul & Baskaya, Feza (2010) Using syllables as indexing terms in full-text information retrieval. In Inguna Skadina & Andrejs Vasiljevs (eds.), *Human Language Technologies: The Baltic Perspective*, pp. 225–232. Amsterdam: IOS Press.
Köhler, Reinhard (1984) Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika* 6: 177–183.
Kulacka, Agnieszka (2010) The coefficients in the formula for the Menzerath-Altmann law. *Journal of Quantitative Linguistics* 17: 257–268.
Menzerath, Paul (1954) *Die Architektonik des deutschen Wortschatzes*. Bonn: Dümmler.
Milicka, Jiri (2014) Menzerath's law: The whole is greater than the sum of its parts. *Journal of Quantitative Linguistics* 21: 85–99.
Szmrecsányi, Benedikt (2004) On operationalizing syntactic complexity. *Annals of the 7th International Conference of Statistical Text Analysis (JADT)*, 1031–1038.
Teupenhayn, Regina & Altmann, Gabriel (1984) Clause length and Menzerath's law. *Glottometrika* 6: 127–138.

Data sources

Arvaniti, Amalia (1999) Standard Modern Greek. *Journal of the International Phonetic Association* 29: 167–172.
Breen, Gavan & Dobson, Veronica (2005) Central Arrernte. *Journal of the International Phonetic Association* 35: 249–254.
Clynes, Adrian & Deterding, David (2011) Standard Malay (Brunei). *Journal of the International Phonetic Association* 41: 259–268.
Cruz-Ferreira, Madalena (1999) Portuguese (European). In *IPA*, pp. 126–130.
Dawd, Abushush & Hayward, Richard (2002) Nara. *Journal of the International Phonetic Association* 32: 249–255.

DiCanio, Christian (2010) Itunyoso Trique. *Journal of the International Phonetic Association* 40: 227–238.
Eaton, Helen (2006) Sandawe. *Journal of the International Phonetic Association* 36: 235–242.
Fougeron, Cécile & Smith, Caroline (1999) French. In *IPA*, pp. 78–81.
Gordon, Matthew; Munro, Pamela & Ladefoged, Peter (2001) Chickasaw. *Journal of the International Phonetic Association* 31: 287–290.
Hamann, Silke & Kula, Nancy (2015) Bemba. *Journal of the International Phonetic Association* 45: 61–69.
Hargus, Sharon & Beavert, Virginia (2014) Northwest Sahaptin. *Journal of the International Phonetic Association* 44: 320–342.
Hayward, Katrina & Hayward, Richard (1999) Amharic. In *IPA*, pp. 45–50.
Hualde, José; Lujanbio, Oihana & Zubiri, Juan (2010) Goizueta Basque. *Journal of the International Phonetic Association* 40: 113–127.
Ido, Shinji (2014) Bukharan Tajik. *Journal of the International Phonetic Association* 44: 87–102.
Ikekeonwu, Clara (1999) Igbo. In *IPA*, pp. 108–110.
Kahn, Sameer (2010) Bengali (Bangladeshi Standard). *Journal of the International Phonetic Association* 40: 221–225.
Kanu, Sullay & Tucker, Benjamin (2010) Temne. *Journal of the International Phonetic Association* 40: 247–253.
Keane, Elinor (2004) Tamil. *Journal of the International Phonetic Association* 34: 111–116.
Khatiwada, Rajesh (2009) Nepali. *Journal of the International Phonetic Association* 39: 373–380.
Kirby, James (2011) Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41: 381–392.
Kohler, Klaus (1999) German. In *IPA*, pp. 86–89.
Laufer, Asher (1999) Hebrew. In *IPA*, pp. 96–99.
Lee, Hyun Bok (1999) Korean. In *IPA*, pp.120–123.
Lee, Wai-Sum & Zee, Eric (2003) Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33: 109–112.
Majidi, Mohammad & Ternes, Elmar (1999) Persian (Farsi). In *IPA*, pp. 124–125.
Marlett, Stephen; Moreno, Xavier & Herrera, Genaro (2005) Seri. *Journal of the International Phonetic Association* 35: 117–121.
Martínez, Eugenio; Fernández, Ana & Carrera, Josefina (2003) Castilian Spanish. *Journal of the International Phonetic Association* 33: 255–260.
Masaquiza, Fanny & Marlett, Stephen (2008) Salasaca Quichua. *Journal of the International Phonetic Association* 38: 223–227.
Ní Chasaide, Ailbhe (1999) Irish. In *IPA*, pp. 111–116.
Ohala, Manjari (1999) Hindi. In *IPA*, pp. 100–103.
Okada, Hideo (1999) Japanese. In *IPA*, pp. 117–119.
Padayodi, Cécile (2008) Kabiye. *Journal of the International Phonetic Association* 38: 215–221.

Pickett, Velma; Villalobos, María & Marlett, Stephen (2010) Isthmus (Juchitán) Zapotec. *Journal of the International Phonetic Association* 40: 365–372.
Remijsen, Bert & Manyang, Caguor (2009) Luanyjang Dinka. *Journal of the International Phonetic Association* 39: 123–124.
Roach, Peter (2004) British English: Received Pronunciation. *Journal of the International Phonetic Association* 34: 239–245.
Ridouane, Rachid (2014) Tashlhiyt Berber. *Journal of the International Phonetic Association* 44: 207–221.
Sadowsky, Scott; Painequeo, Héctor; Salamanca, Gastón & Avelino, Heriberto (2013) Mapudungun. *Journal of the International Phonetic Association* 43: 87–96.
Schuh, Russell & Yalwa, Lawan (1999) Hausa. In *IPA*, pp. 90–95.
Shosted, Ryan & Chikovani, Vakhtang (2006) Standard Georgian. *Journal of the International Phonetic Association* 36: 255–264.
Soderberg, Craig; Ashley, Seymour & Olson, Kenneth (2012) Tausug (Suluk). *Journal of the International Phonetic Association* 42: 361–364.
Szende, Tamás (1999) Hungarian. In *IPA*, pp. 104–107.
Thelwall, Robin & Sa'adeddin, Akram (1999) Arabic. In *IPA*, pp. 51–54.
Tingsabadh, Kalaya & Abramson, Arthur (1999) Thai. In *IPA*, pp. 147–150.
Tuttle, Siri & Sandoval, Merton (2002) Jicarilla Apache. *Journal of the International Phonetic Association* 32: 105–112.
Urquía, Rittma & Marlett, Stephen (2008) Yine. *Journal of the International Phonetic Association* 38: 365–369.
Valenzuela, Pilar & Gussenhoven, Carlos (2013) Shiwilu (Jebero). *Journal of the International Phonetic Association* 43: 97–106.
Watkins, Justin (2001) Burmese. *Journal of the International Phonetic Association* 31: 291–295.
Yanushevskaya, Irena & Buncic, Daniel (2015) Russian. *Journal of the International Phonetic Association* 45: 221–228.
Zee, Eric (1999) Chinese (Hong Kong Cantonese). In *IPA*, pp. 58–60.
Zimmer, Karl & Orgun, Orhan (1999) Turkish. In *IPA*, pp. 154–156.

Appendix 1. Data from "The North Wind and the Sun" dataset

Language	Region	Family	Phonemes	Words	Clauses	Phon/ Word	Word/ Clause
Amharic	Africa	Afro-Asiatic	661	94	8	7.03	11.75
Apache	America	Na-Dene	579	118	15	4.91	7.87
Arabic	West Asia	Afro-Asiatic	488	85	9	5.74	9.44
Arrernte	East Asia	Pama-Nyungan	436	73	12	5.97	6.08
Basque	Europe	Vasconic	401	83	7	4.83	11.86
Bemba	Africa	Niger-Congo	435	79	8	5.51	9.88
Bengali	West Asia	Indo-European	459	104	10	4.41	10.40
Burmese	East Asia	Sino-Tibetan	300	42	7	7.14	6.00
Cantonese	East Asia	Sino-Tibetan	351	91	10	3.86	9.10

Language	Region	Family	Phonemes	Words	Clauses	Phon/ Word	Word/ Clause
Chickasaw	America	Muskogean	474	57	10	8.32	5.70
Dinka	Africa	Nilo-Saharan	548	137	10	4.00	13.70
English	Europe	Indo-European	383	113	9	3.39	12.56
French	Europe	Indo-European	343	108	9	3.18	12.00
Georgian	West Asia	Caucasian	418	70	9	5.97	7.78
German	Europe	Indo-European	452	108	10	4.19	10.80
Greek	Europe	Indo-European	479	104	9	4.61	11.56
Hausa	Africa	Afro-Asiatic	648	166	12	3.90	13.83
Hebrew	West Asia	Afro-Asiatic	526	89	11	5.91	8.09
Hindi	West Asia	Indo-European	467	125	8	3.74	15.63
Hungarian	Europe	Uralic	431	100	10	4.31	10.00
Igbo	Africa	Niger-Congo	356	107	8	3.33	13.38
Irish	Europe	Indo-European	406	129	7	3.15	18.43
Japanese	East Asia	Japonic	444	89	9	4.99	9.89
Kabiye	Africa	Niger-Congo	433	91	9	4.76	10.11
Korean	East Asia	Koreanic	381	60	7	6.35	8.57
Malay	East Asia	Austronesian	481	78	8	6.17	9.75
Mandarin	East Asia	Sino-Tibetan	421	98	10	4.30	9.80
Mapudungun	America	Araucanian	360	75	9	4.80	8.33
Nara	Africa	Nilo-Saharan	466	108	11	4.31	9.82
Nepali	West Asia	Indo-European	502	95	9	5.28	10.56
Persian	West Asia	Indo-European	483	91	9	5.31	10.11
Portuguese	Europe	Indo-European	380	98	8	3.88	12.25
Quichua	America	Quechuan	593	94	11	6.31	8.55
Russian	Europe	Indo-European	468	97	9	4.82	10.78
Sahaptin	America	Penutian	375	57	8	6.58	7.13
Sandawe	Africa	Khoisan	383	79	9	4.85	8.78
Seri	America	Hokan	593	157	11	3.78	14.27
Shiwilu	America	Kawapanan	837	108	14	7.75	7.71
Spanish	Europe	Indo-European	425	97	9	4.38	10.78
Tajik	West Asia	Indo-European	482	90	7	5.36	12.86
Tamil	West Asia	Dravidian	541	79	9	6.85	8.78
Tashlhiyt	Africa	Afro-Asiatic	306	76	9	4.03	8.44
Tausug	East Asia	Austronesian	572	114	12	5.02	9.50
Temne	Africa	Niger-Congo	446	125	11	3.57	11.36
Thai	East Asia	Tai-Kadai	480	131	11	3.66	11.91
Trique	America	Oto-Manguean	359	107	10	3.36	10.70
Turkish	West Asia	Turkic	431	66	9	6.53	7.33
Vietnamese	East Asia	Austro-Asiatic	334	117	7	2.85	16.71
Yine	America	Arawakan	559	63	10	8.87	6.30
Zapotec	America	Oto-Manguean	327	87	9	3.76	9.67
Average			**458.06**	**96.18**	**9.44**	**4.76**	**10.19**

Appendix 2. Instrumental variables

Language	Region	Consonants	Vowels	Tones	Cases	Genders	Inflections
Amharic	Africa	27	7	1	2	2	6
Apache	America	33	8	3	1	1	5
Arabic	West Asia	29	6	1	1	2	6
Arrernte	East Asia	27	4	1	8	1	4
Basque	Europe	23	5	1	10	1	4
Bemba	Africa	26	10	2	1	5	4
Bengali	West Asia	29	7	1	6	2	2
Burmese	East Asia	34	9	4	8	1	2
Cantonese	East Asia	19	11	6	1	1	1
Chickasaw	America	16	9	1	2	1	6
Dinka	Africa	20	7	4	1	1	6
English	Europe	24	11	1	2	1	2
French	Europe	20	13	1	1	2	4
Georgian	West Asia	28	5	1	6	1	8
German	Europe	23	15	1	4	3	2
Greek	Europe	18	5	1	3	3	4
Hausa	Africa	28	10	2	1	2	6
Hebrew	West Asia	25	5	1	1	2	4
Hindi	West Asia	34	11	1	2	2	2
Hungarian	Europe	25	14	1	10	1	4
Igbo	Africa	26	8	3	1	1	6
Irish	Europe	35	11	1	2	2	2
Japanese	East Asia	16	5	2	8	1	4
Kabiye	Africa	21	9	2	1	1	2
Korean	East Asia	19	18	1	6	1	6
Malay	East Asia	18	6	1	1	1	4
Mandarin	East Asia	19	6	4	1	1	1
Mapudungun	America	22	6	1	2	1	8
Nara	Africa	25	10	2	5	2	4
Nepali	West Asia	27	11	1	2	1	4
Persian	West Asia	23	6	1	2	1	4
Portuguese	Europe	19	13	1	1	2	4
Quichua	America	23	3	1	8	1	8
Russian	Europe	36	6	1	6	3	4
Sahaptin	America	32	7	1	4	1	10
Sandawe	Africa	44	15	2	1	5	8
Seri	America	18	8	1	1	1	5
Shiwilu	America	17	4	1	6	1	6
Spanish	Europe	19	5	1	1	2	4
Tajik	West Asia	22	6	1	2	1	4

Language	Region	Consonants	Vowels	Tones	Cases	Genders	Inflections
Tamil	West Asia	15	10	1	6	3	2
Tashlhiyt	Africa	34	3	1	2	2	6
Tausug	East Asia	17	3	1	1	1	4
Temne	Africa	19	9	2	1	5	2
Thai	East Asia	21	9	5	1	1	2
Trique	America	29	8	9	1	1	6
Turkish	West Asia	22	8	1	6	1	6
Vietnamese	East Asia	22	11	8	1	1	1
Yine	America	16	5	1	2	4	6
Zapotec	America	20	5	3	1	1	8
Average		**24.08**	**8.12**	**1.92**	**3.08**	**1.70**	**4.46**

Contact information:

Germán Coloma
CEMA University
Av. Córdoba 374
Buenos Aires, C1054AAP
Argentina
e-mail: gcoloma(at)cema(dot)edu(dot)ar