Germán Coloma*

The existence of negative correlation between linguistic measures across languages

DOI 10.1515/cllt-2015-0020

Abstract: This paper proposes a procedure to evaluate the possible existence of negative correlation between three language ratios (phonemes per syllable, syllables per word, and words per clause), based on a synergetic linguistics' approach. It uses partial correlation coefficients and simultaneous-equation regressions, and the analysis is performed on data obtained from the fable "The North Wind and the Sun", translated into 50 languages. After controlling for phenomena related to geographic and genetic factors, we end up with the conclusion that the three ratios are negatively correlated between themselves, and this can be seen as a signal of the possible existence of complexity trade-offs.

Keywords: language ratios, partial correlation, simultaneous equations, complexity trade-off, synergetic linguistics

1 Introduction

The aim of this paper is to propose a procedure to evaluate the possible existence of negative correlation between different linguistic measures. The procedure is based on simultaneous-equation regressions, which is a statistical technique that is widely used in other social sciences. The analysis is performed using a cross-linguistic perspective, and the measures used are three language ratios: phonemes per syllable, syllables per word and words per clause.

The three abovementioned ratios can be seen as indicators of different levels of language complexity. This concept is linked to the idea that a language becomes more complex when "the number of parts in a system" increases,[1] and it assumes that complexity is measurable using indicators for different language dimensions (e. g., phonology, morphology, syntax).

[1] This idea of complexity is sometimes referred to as "absolute complexity", as opposed to "relative complexity" (which is a concept related to the difficulty of processing, acquisition and learning). In this topic, see Vulanovic (2007) and Miestamo (2008).

***Corresponding author: Germán Coloma,** CEMA University, Buenos Aires, Argentina,
E-mail: gcoloma@cema.edu.ar

A complexity trade-off is a situation in which a higher level of complexity for a certain language component appears in correspondence to a lower level of complexity for another component. As each of the ratios that we consider can be interpreted as a measure that represents a different language domain, then a negative relationship between two of those ratios can be read as a signal of a complexity trade-off between two different language components.

In order to evaluate the relationships between phonemes per syllable, syllables per word and words per clause, we use a single text available in several languages. That text is the fable known as "The North Wind and the Sun", which is a story recommended by the International Phonetic Association (IPA) to illustrate the phonetics of different languages and dialects. Having that text for a considerable number of languages, we have selected a sample of 50 cases and calculated the corresponding ratios.

The simplest way to analyze the existence of correlation in a context like the one described is to calculate standard (Pearson) correlation coefficients between the different ratios. We also calculate the so-called "partial correlation coefficients", which have the advantage of using information about the three ratios at the same time. Alternatively, we run regressions in which each language ratio is supposed to depend on other variables (e. g., the other ratios, plus some categorical variables related to geographic and genetic factors). Those regressions are run at the same time, as a system of simultaneous equations, and their results are used to compute new partial correlation coefficients. Those coefficients are compared among themselves, and they are related to previous results that appear in the literature.

The rest of this article is organized as follows. In Section 2 we review the literature about cross-linguistic analyses of correlation, especially the one in which we find results about possible relationships between different measures of complexity. In Section 3 we explain the main characteristics of the dataset that we have assembled, based on the transcriptions for "The North Wind and the Sun" that appear in IPA (1999) and other related sources. In Section 4 we provide a theoretical foundation for our analysis, based on a synergetic linguistics' approach. In Section 5 we apply regression techniques to our dataset, and find some results that can be compared with the ones from the existing literature. Finally, in Section 6 we make some concluding remarks.

2 Review of the literature

The literature about quantitative analysis of cross-linguistic correlations is not particularly large, but it is considerably diverse. The branch of that literature which is closer to the type of analysis that we perform in this article is the one that began with Fenk-Oczlon and Fenk (1985). That literature analyzes the possible existence of correlation between different language sub-systems using actual text analysis, and it does so by calculating several linguistic ratios (phonemes per syllable, syllables per word, phonemes per word, words per clause, etc.). In Fenk-Oczlon and Fenk (1993), for example, we find a correlation analysis of those ratios that comes from translations of a number of simple German sentences into 33 different languages, while in Fenk-Oczlon and Fenk (2005) there is an additional analysis in which some of those ratios are correlated with theoretical variables (e. g., number of cases, tendency to postpositions, order of object and verb). They end up with a set of conclusions that imply negative correlation between many of the analyzed variables, and some of them turn out to be statistically significant.

Another branch of the literature, which leads to results that are similar to Fenk-Oczlon and Fenk's, is the one that relates word length to measures of phonological complexity. This literature begins with a paper by Nettle (1995), which finds a strong negative correlation between phoneme inventory size and phonemes per word, and is later continued by studies such as Wichmann et al. (2011), which uses an alternative measure of phonological complexity and a much larger dataset (while also controlling for geographic and genetic factors). Another contribution that is worth noting is the one by Oh et al. (2013), which calculates correlations between syllables per word and two measures of "information density" (which are ratios between the number of syllables and the number of words in the same text translated to different languages, using Vietnamese as an external reference). These authors find a highly significant negative correlation between syllabic information density and number of syllables per word.

Other articles, such as Shosted (2006), study the possible existence of negative correlation between linguistic phenomena using purely theoretical (or "typological") variables. In that paper, the author compares the number of possible syllable types in a sample of 32 languages with the number of inflectional categories of the verbs in those languages, and his analysis ends up with the conclusion that no statistically significant correlation can be found. In the same line of research, Nichols (2009) analyzes five language domains (phonology, inflectional synthesis, classification, syntax and lexicon), using a sample of 68 languages. While she finds no significant negative correlation between any of those five domains, she does find a significant positive correlation between her measure of morphological synthesis and her measure of syntactic structure.

In many cases, this literature considers the measures that it uses as indicators of language complexity, and therefore interprets its correlation coefficients as evidence for or against the existence of complexity trade-offs. Part of the quantitative linguistic literature has also found phenomena related to negative correlation between different dimensions of actual text complexity using other statistical tools, such as spectral analysis. Moscoso (2011), for example, has detected the existence of a certain “universal shape” of languages related to the absence of repetition in very short scales and the abundance of repetition in very long scales, with a relatively flat regime for intermediate scales. As those scales can be loosely related to the concepts of word, sentence and discourse, those relationships (that are studied for relatively large corpora and applied to a sample of 15 languages) reveal a negative correlation between morphological and syntactic structure.

Other approaches to this problem restrict themselves to a particular language subsystem. Maddieson (2007), for example, measures phonological complexity using three different indicators: inventory size (of both consonants and vowels), tone systems, and elaboration of the syllable inventory. He also looks for relationships between those measures, but he only finds a significant positive correlation between consonant inventory and syllable structure, and a (less significant) negative correlation between syllable structure and tone complexity. Moran and Blasi (2014), conversely, find a relatively large negative correlation between vowel inventory size and number of phonemes per word, in a study in which this last concept is measured using the same word lists used by Wichmann et al. (2011).

In this paper, we will focus on the correlation between three linguistic ratios (phonemes per syllable, syllables per word, and words per clause) using empirical measures derived from actual texts, and those measures will be subject to statistical analysis to determine the possible existence of negative correlation between them. As each of these ratios can be seen as an indicator of a certain type of complexity, their correlations also have an interpretation related to the existence of complexity trade-offs.

Another important feature of this paper has to do with its level of statistical sophistication. Contrary to previous studies that restrict themselves to product–moment correlation coefficients, we also use partial correlation coefficients calculated in different ways. With that approach, we control for factors such as the interaction of the proposed indicators and the effect of geographic and genetic factors, in a context of simultaneous-equation system estimations.

3 The North Wind and the Sun

The fable of the North Wind and the Sun, attributed to Aesop, is a text that has been used for many decades by the International Phonetic Association as a “specimen” or model to illustrate the sounds of languages, and also the phonetic symbols that are suitable to describe those sounds. It is therefore a unique case of a short text for which specialists in the phonetics of different languages have analyzed the sounds, the phonemes, the syllables and the words of the languages and dialects under study. For example, the (Standard Southern British) English version of "The North Wind and the Sun" that appears in Roach (2004) is the following:

> *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.*

and its corresponding phonemic transcription is this:

> ðə nɔ:θ wɪnd ən ðə sʌn wə dɪspju:tɪŋ wɪtʃ wəz ðə strɒŋgə | wən ə trævlə keɪm əlɒŋ ræpt ɪn ə wɔ:m kləʊk || ðəɪ əgri:d ðət ðə wʌn hu fɜ:st səksi:dɪd ɪn meɪkɪŋ ðə trævlə teɪk hɪz kləʊk ɒf ʃʊd bi kənsɪdəd strɒŋgə ðən ði ʌðə || ðən ðə nɔ:θ wɪnd blu: əz hɑ:d əz i kʊd | bət ðə mɔ: hi blu: ðə mɔ: kləʊsli dɪd ðə trævlə fəʊld hɪz kləʊk əraʊnd hɪm | ænd ət lɑ:st ðə nɔ:θ wɪnd geɪv ʌp ði ətempt || ðən ðə sʌn ʃɒn aʊt wɔ:mli | ænd əmi:diətli ðə trævlə tʊk ɒf ɪz kləʊk || n səʊ ðə nɔ:θ wɪn wəz əblaɪdʒd tu kənfes ðət ðə sʌn wəz ðə strɒŋgr əv ðə tu: ||

If we count the number of clauses, words, syllables and phonemes in this text, we find that it has 9 clauses,[2] 113 words, 143 syllables and 383 phonemes. This allows to calculate the average number of phonemes per syllable (equal to 2.6783), syllables per word (equal to 1.2655), and words per clause (equal to 12.56). The same exercise can be performed for other available languages. For example, the Japanese version of "The North Wind and the Sun" (Okada 1999) has a lower number of phonemes per syllable (equal to 1.9559) but a higher number of syllables per word (equal to 2.5506). That figure is even larger in the Tamil version (Keane 2004), where there are 3.1899 syllables per word but only 8.78 words per clause.

The systematic computation of the phoneme/syllable, syllable/word and word/clause ratios for different versions of "The North Wind and the Sun" can be a useful source to compare the value of those measures across languages, and to capture the possible existence of correlation between them. In order to perform that analysis, we selected a sample of 50 languages for which we found versions of the abovementioned text in either the *Handbook of the International*

[2] 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.

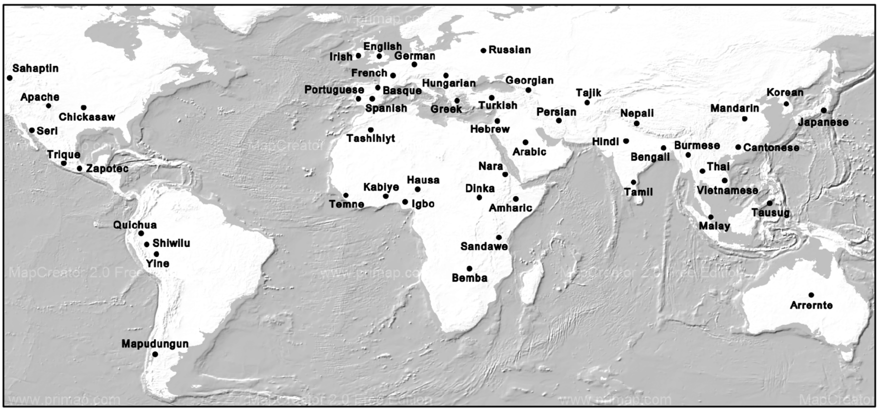

Figure 1: Location of the languages included in the sample.

Phonetic Association (IPA 1999) or in the series of "Illustrations of the IPA", published by the *Journal of the International Phonetic Association* (see Figure 1). The sample assembled consists of ten languages from each of the five areas in which we divided the world: 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, Temne, Kabiye, Igbo, Hausa, Dinka, Nara, Amharic, Sandawe and Bemba), West Asia (Georgian, Turkish, Hebrew, Arabic, Persian, Tajik, Nepali, Hindi, Bengali and Tamil) and East Asia (Japanese, Korean, Mandarin, Cantonese, Burmese, Thai, Vietnamese, Malay, Tausug and Arrernte).[3] This sample is basically the same one that we used in another paper about a different topic (Coloma 2015), and covers 26 different language families.

Table 1 shows the main descriptive statistics of our dataset for the calculated linguistic ratios, in terms of their maximum, minimum and average values, together with their standard deviations (both at an aggregate level and within each of the areas).[4] In it, we can observe that, on average, the number of phonemes per syllable in the whole dataset is 2.3004, but that ratio ranges from a minimum of 1.7115 (which corresponds to Igbo, a Niger-Congo language spoken in Nigeria) to a maximum of 2.8547 (which corresponds to Vietnamese).[5]

[3] In this division, Australia (where the Arrernte language is spoken) is considered as a part of East Asia.

[4] For the complete set of linguistic ratios, see Appendix 1.

[5] In order to define the number of phonemes in each version of "The North Wind and the Sun", we tried to follow the assumptions used by the authors who wrote the corresponding illustrations of the IPA, but we also applied some unifying criteria. For example, long, short, oral and

Table 1: Descriptive statistics for "The North Wind and the Sun" dataset.

Concept	Maximum	Minimum	Average	Std. Dev.
Phonemes/Syllables	**2.8547**	**1.7115**	**2.3004**	**0.2375**
America	2.5761	1.9617	2.3028	0.1911
Europe	2.6783	2.0430	2.2922	0.2033
Africa	2.5521	1.7115	2.1466	0.2763
West Asia	2.5288	2.1468	2.3079	0.1292
East Asia	2.8547	1.9559	2.4525	0.2896
Syllables/Words	**3.7460**	**1.0000**	**2.1869**	**0.5996**
America	3.7460	1.5414	2.5112	0.7548
Europe	2.2530	1.2655	1.7907	0.3300
Africa	2.8481	1.6053	2.1185	0.4508
West Asia	3.1899	1.6640	2.3937	0.4319
East Asia	3.1190	1.0000	2.1201	0.7348
Words/Clauses	**18.4286**	**5.7000**	**10.3313**	**2.6824**
America	14.2727	5.7000	8.6224	2.4719
Europe	18.4286	10.0000	12.1002	2.3627
Africa	13.8333	8.4444	11.1048	2.0129
West Asia	15.6250	7.3333	10.0973	2.5317
East Asia	16.7143	6.0000	9.7317	3.0288

Vietnamese is also the language with the lowest syllable/word ratio (equal to 1, since in the Vietnamese version of "The North Wind and the Sun" all the words are monosyllabic), while the language with the highest syllable/word ratio (equal to 3.7460) is Yine (an Arawakan language spoken in Peru), and the overall average number of syllables per word is 2.1869. Finally, the minimum value for the word/clause ratio is 5.7 and corresponds to Chickasaw (Muskogean, spoken in the United States), while the maximum value for that ratio is 18.4286 (Irish), in a context where the average number of words per clause is 10.3313.[6]

The description of the variables shown in Table 1 can be complemented with the calculation of product–moment correlation coefficients. Those coefficients

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 counted as a combination of two (identical) consecutive phonemes.

[6] These calculations are based on the criteria used by the authors of the different illustrations of the IPA to divide the texts into words. These criteria may not be universally accepted, especially when we deal with some East Asian languages in which the distinction between syllables and words is not so clear. I thank Justin Watkins for this observation.

Table 2: Correlation coefficients for “The North Wind and the Sun” dataset.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000	−0.2420	−0.0522
Syllables per word		1.0000	−0.6785
Words per clause			1.0000

appear in Table 2, in which the highest absolute value is the one that corresponds to the correlation between the syllable/word ratio and the word/clause ratio, which is equal to −0.6785. The correlation coefficient between phonemes per syllable and syllables per word, conversely, is much smaller ($r = -0.2420$), while the correlation coefficient between phonemes per syllable and the words per clause is even smaller ($r = -0.0522$).

The absolute values of those coefficients are related to their statistical significance. For example, correlation between syllables per word and words per clause is significantly different from zero at a 1% probability level ($p = 0.0000$), while correlation between phonemes per syllable and syllables per word is only statistically significant at a 10% probability level ($p = 0.0905$), and correlation between phonemes per syllable and words per clause fails to be significant at a any reasonable probability level ($p = 0.7187$).

4 Theoretical foundation

The existence of negative correlation between linguistic measures such as the phoneme/syllable, syllable/word and word/clause ratios can be related to different processes through which languages trade off complexity in one dimension with simplicity in another one. In this section we provide a theoretical justification for those complexity trade-offs, based on an approach known as “synergetic linguistics”. The idea behind this approach is that language is “a self-organizing and self-regulating system whose properties come from the interaction of several constitutive, forming and control requirements” (Köhler 2005). Among them we can mention the so-called “coding requirement” (i. e., the need to provide expressions for given meanings), the “economy requirement” (i. e., the need to use as few elements as possible to produce the desired expressions), and the “stability requirement” (i. e., the need to modify language as little as possible so that it could be understood).

Synergetic linguistics’ models are sometimes based on the idea that language systems are a solution to an optimization problem whose aim is to maximize or to minimize a particular function or combination of functions.[7] Köhler (1987), for example, illustrates this idea through a case in which language is assumed to minimize “production effort”, “memory effort” and “decoding effort”, using word length, phoneme inventory size and degree of similarity as control variables. In the problem under analysis in this paper, we can apply a similar model, using our three linguistic ratios (phonemes per syllable, syllables per word and words per clause) as the corresponding control variables.

Let us assume that the three ratios under analysis contribute to satisfy the coding requirement of a language, since larger units in any of those dimensions (i. e., syllables with more phonemes, or words with more syllables, or clauses with more words) are helpful to provide expressions for the meanings that languages have to convey. Increasing complexity in any of those dimensions, however, is bad in terms of the cost that one has to incur to produce the required expressions (production effort) or to remember those expressions (memory effort). Long syllables, words and clauses are therefore inconvenient to satisfy the economy requirement of a given language, but any language can combine shorter and longer components so as to maximize a function that increases in decoding ease and decreases in production and memory effort.

In Figure 2 we have drawn a diagram in which we illustrate that combination, assuming that our three variables make decoding easier (maxD), but phonemes per syllable (Phon/Syll) and syllables per word (Syll/Word) are costly in terms of production effort (minP), and syllables per word and words per clause (Word/Clause) are costly in terms of memory effort (minM).

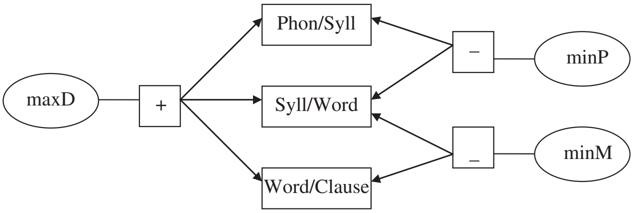

Figure 2: Diagram of a language system.

The value that these three ratios take in any particular language can be determined by a process in which each of them helps to increase decoding ease but has a cost in terms of production or memory effort. When a language uses one of these variables in a more complex level (i. e., it has a larger value for the

[7] For an explanation of this, see Altmann and Wimmer (2007).

phoneme/syllable, or the syllable/word, or the word/clause ratios), it should compensate that complexity by having a greater simplicity in another dimension. This is the main theoretical basis for complexity trade-offs, which in the context that we are studying should be reflected in the existence of negative correlation between the linguistic measures under analysis.

That correlation, moreover, is expected to be stronger for measures that are directly related (e. g., phonemes per syllable versus syllables per word, and syllables per word versus words per clause) and weaker for measures that are less related (i. e., phonemes per syllable versus words per clause). This can also be linked to the fact that syllables are the constituents of words, while words are the constituents of clauses. If a language has longer syllables, then those syllables will be able to convey more information, and it will therefore be possible to have shorter words. Similarly, if a language has longer words, that language will be able to use fewer words per clause (because each word can have multiple morphemes, and those morphemes can play the same roles that, in other languages, are played by individual words).

5 Statistical analysis

5.1 Correlation and regression coefficients

In the last part of Section 3, we included a table that shows the product–moment correlation coefficients between phonemes per syllable, syllables per word and words per clause. Another way to appreciate the relationships between those variables is to plot each observation in a graph that represents two of those measures at the same time. This is what we do in Figures 3–5, which correspond to phonemes per syllable versus syllables per word, syllables per word versus words per clause, and phonemes per syllable versus words per clause.

In each of those figures, we have also depicted two straight lines. They represent the results of running linear regressions to relate the corresponding variables. The full line corresponds to the regression for which the independent variable is the one that appears on the abscissa axis of the graph, while the dotted line corresponds to the regression for which the independent variable is the one that appears on the ordinate axis.

Graphically speaking, one can think of correlation as the inverse of the average distance between the two lines drawn in each figure. On Figure 4, for example, that distance is relatively small, and therefore the correlation

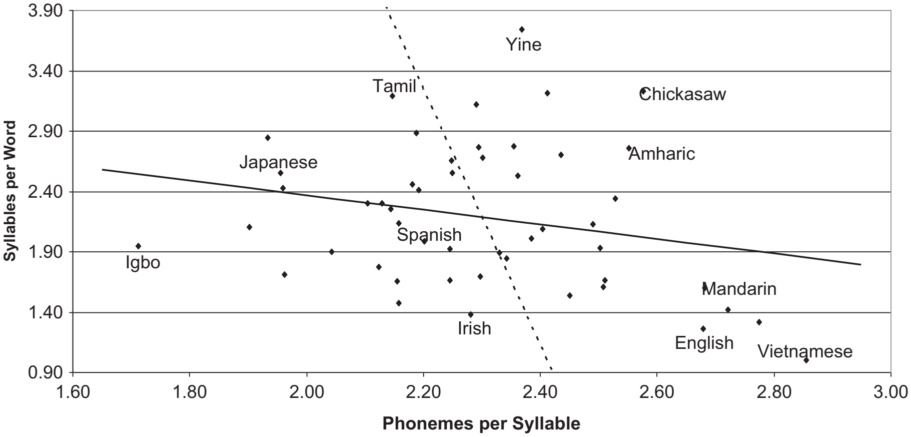

Figure 3: Phonemes per syllable versus syllables per word.

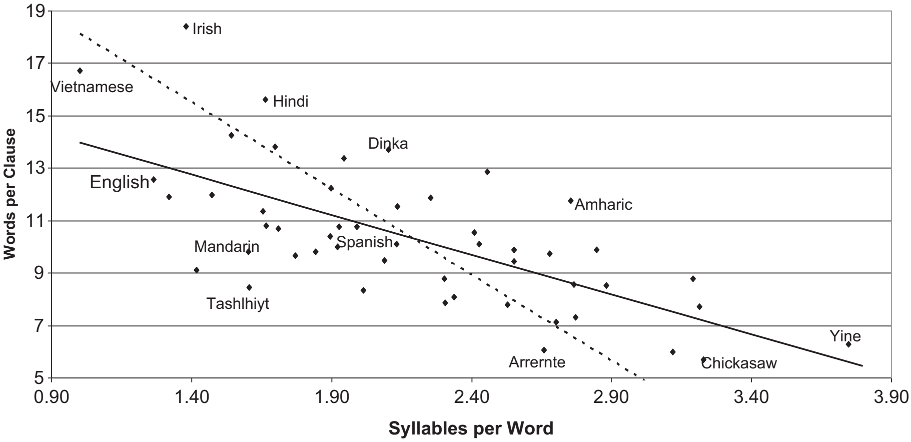

Figure 4: Syllables per word versus words per clause.

coefficient between the two variables ($r = -0.6785$) is considerably large in absolute value. On the contrary, on Figure 5 the distance between the two lines is very large, and this is related to the fact that the corresponding correlation coefficient ($r = -0.0522$) is small.

The product–moment correlation coefficients also have a relationship with the slopes of the lines represented in the figures. As all those coefficients are negative, the depicted lines are negatively sloped. But the most exact

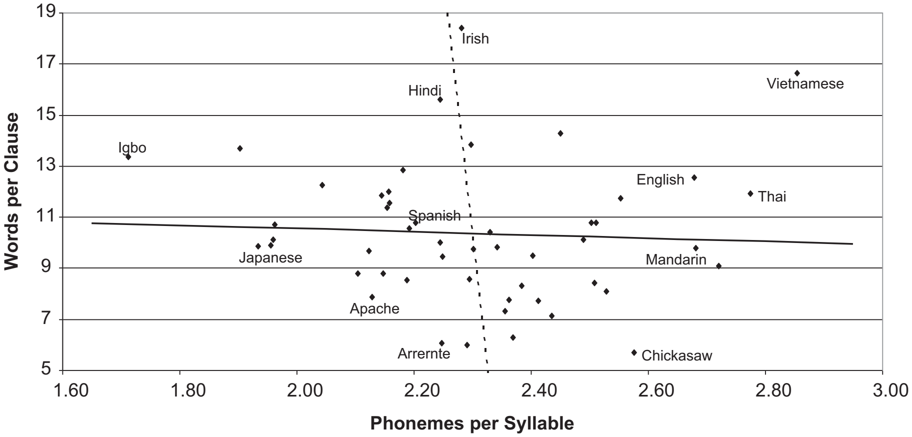

Figure 5: Phonemes per syllable versus words per clause.

relationship between the correlation coefficients calculated in Section 3 and the regression lines represented in Figures 3–5 appears if we take into account the values of the slopes of those regression lines, and their corresponding goodness-of-fit measures.

Let us consider, for example, the results obtained when one regresses *Phon/Syll* against *Syll/Word*, and *Syll/Word* against *Phon/Syll*. These are the following:

$$Phon/Syll = 2.509973 - 0.095826.Syll/Word;\ R^2 = 0.05855 \qquad [1];$$

$$Syll/Word = 3.573388 - 0.604350.Phon/Syll;\quad R^2 = 0.05855 \qquad [2].$$

As we can see, both regressions produce the same R^2 coefficient (R^2 = 0.05855), and the value of that coefficient is exactly the square of the correlation coefficient between *Phon/Syll* against *Syll/Word* (r = −0.2420). Moreover, if we look at the regression coefficients that correspond to the slopes of those regression lines ("β_1 = −0.095826" and "β_2 = −0.604350") and find the square root of the product of those coefficients, we obtain the same result, which is equal to the correlation coefficient between *Phon/Syll* against *Syll/Word*. This implies performing the following calculation:

$$r(PS, SW) = -\sqrt{\beta_1 \cdot \beta_2} = -\sqrt{(-0.095826) \cdot (-0.604350)} = -0.2420 \qquad [3];$$

where *PS* stands for "phonemes per syllable" and *SW* stands for "syllables per word". Note that in this formula we assume that, as both regression coefficients are negative, the corresponding product–moment correlation coefficient must also be negative.

The same situation occurs when we run regressions between *Syll/Word* and *Word/Clause*, *Word/Clause* and *Syll/Word*, *Phon/Syll* and *Word/Clause*, and *Word/Clause* and *Phon/Syll*. Their results are the following:

$$Syll/Word = 3.753594 - 0.151650 \cdot Word/Clause; \quad R^2 = 0.46031 \qquad [4];$$

$$Word/Clause = 16.96909 - 3.035319 \cdot Syll/Word; \quad R^2 = 0.46031 \qquad [5];$$

$$Phon/Syll = 2.348172 - 0.004623 \cdot Word/Clause; \quad R^2 = 0.00273 \qquad [6];$$

$$Word/Clause = 11.68831 - 0.589895 \cdot Phon/Syll; \quad R^2 = 0.00273 \qquad [7];$$

and the corresponding correlation coefficients can be calculated as:

$$r(SW, WC) = -\sqrt{\beta_3 \cdot \beta_4} = -\sqrt{(-0.151650) \cdot (-3.035319)} = -0.6785 \qquad [8];$$

$$r(PS, WC) = -\sqrt{\beta_5 \cdot \beta_6} = -\sqrt{(-0.004623) \cdot (-0.589895)} = -0.0522 \qquad [9];$$

where *WC* stands for “words per clause”.

5.2 Partial correlation coefficients

The product–moment correlation coefficients are measures that represent relationships between different variables, but they do not take into account any possible interdependence between those variables and other outside factors. One further step in the statistical analysis of the data is to consider that interdependence, and the most direct way to do it is to compute partial correlation coefficients.

A partial correlation coefficient is a measure of the linear dependence for a pair of variables in the case where the influence of other variables is eliminated. To calculate that coefficient, it is necessary to control for the possible effect of other factors on the two variables that we wish to correlate, and to eliminate that effect using some statistical procedure. In our case, having only three variables, a relatively simple strategy to do this consists of using the results obtained when running the regressions corresponding to eqs [1], [2], [4], [5], [6] and [7].

If we compute the residuals of those regressions, the partial correlation coefficients can be calculated as product–moment correlations between different pairs of residuals. For example, the partial correlation coefficient between phonemes per syllable and syllables per word is the correlation coefficient between the residuals from a regression in which phonemes per syllable are regressed against words per clause (eq. [6]), and the residuals from another regression in which syllables per word are regressed against words per clause (eq. [4]).[8]

If we apply this method to calculate the three partial correlation coefficients generated by the relationships between phonemes per syllable, syllables per word and word per clause, we end up with the matrix that appears on Table 3. As we can see, this matrix displays higher absolute values than the ones reported on Table 2. This increase is very important for the correlation coefficient between the phoneme/syllable and syllable/word ratios (which goes from –0.2420 to –0.3781) and even more important for the one that corresponds to the phoneme/syllable and word/clause ratios (which goes from –0.0522 to –0.3036). Both coefficients now become statistically significant at a 5 % probability level, since their corresponding *p*-values are "$p = 0.0074$" and "$p = 0.0340$".

Table 3: Partial correlation coefficients.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000	–0.3781	–0.3036
Syllables per word		1.0000	–0.7132
Words per clause			1.0000

The same coefficients gotten through the correlation of six different regression residuals can be obtained if we apply a single multiple regression procedure. This consists of running three ordinary least-square equations for the following functions:

$$Phon/Syll = c(1) + c(2)^{*}Syll/Word + c(3)^{*}Word/Clause \qquad [10];$$

$$Syll/Word = c(4) + c(5)^{*}Phon/Syll + c(6)^{*}Word/Clause \qquad [11];$$

$$Word/Clause = c(7) + c(8)^{*}Phon/Syll + c(9)^{*}Syll/Word \qquad [12];$$

where *c(1)*, *c(2)*, *c(3)*, *c(4)*, *c(5)*, *c(6)*, *c(7)*, *c(8)* and *c(9)* are the coefficients to be estimated.

With these regression coefficients (whose values are reported on Table 4), the partial correlations between the different linguistic ratios can be calculated by using the same formulae that we applied in eqs [3], [8] and [9]. In this case, this implies to perform the following calculations:

[8] For an explanation of this procedure, and its equivalence with other methods to calculate partial correlation coefficients, see Lowry (2000), Appendix 3a.

Table 4: Regression results to calculate partial correlation coefficients.

Concept	Coefficient	*t*-Statistic	Probability
Phoneme/Syllable equation			
Constant [c(1)]	3.112233	10.33539	0.0000
Syllable/Word [c(2)]	−0.203554	−2.800039	0.0058
Word/Clause [c(3)]	−0.035492	−2.184185	0.0306
R-squared	0.1453		
Syllable/Word equation			
Constant [c(4)]	5.402813	8.512147	0.0000
Phoneme/Syllable [c(5)]	−0.702342	−2.800039	0.0058
Word/Clause [c(6)]	−0.154897	−6.975993	0.0000
R-squared	0.5375		
Word/Clause equation			
Constant [c(7)]	23.48596	7.436737	0.0000
Phoneme/Syllable [c(8)]	−2.596391	−2.184185	0.0306
Syllable/Word [c(9)]	−3.284121	−6.975993	0.0000
R^2	0.5100		

$$r(PS, SW) = -\sqrt{c(2)\cdot c(5)} = -\sqrt{(-0.203554)\cdot(-0.702342)} = -0.3781 \qquad [13];$$

$$r(SW, WC) = -\sqrt{c(6)\cdot c(9)} = -\sqrt{(-0.154897)\cdot(-3.284121)} = -0.7132 \qquad [14];$$

$$r(PS, WC) = -\sqrt{c(3)\cdot c(8)} = -\sqrt{(-0.035492)\cdot(-2.596391)} = -0.3036 \qquad [15];$$

which generate the same partial correlation coefficients reported in Table 3.

5.3 Geographic and genetic factors

The regression coefficients calculated in the previous sections are the result of analyses that assume that each linguistic ratio depends on the other linguistic ratios. In this section, we extend that approach to explore the possibility that the three measures under analysis can also depend on other variables.

Let us suppose, for example, that phonemes per syllable depend on several geographic and genetic factors. One possibility to analyze this is to run a regression in which the phoneme/syllable ratio is the dependent variable, and the independent variables are the syllable/word and word/clause ratios, plus a few categorical variables related to genetic affiliation and geographic location. This could be written in the following way:

$$\begin{aligned} Phon/Syll = {} & c(1)^{*}Europe + c(2)^{*}Africa + c(3)^{*}Westasia + c(4)^{*}Eastasia \\ & + c(5)^{*}America + c(6)^{*}Indoeuro + c(7)^{*}Afroasiatic \\ & + c(8)^{*}Nigercongo + c(9)^{*}Sinotibetan \\ & + c(10)^{*}Syll/Word + c(11)^{*}Word/Clause \end{aligned} \qquad [16];$$

where *Europe*, *Africa*, *Westasia*, *Eastasia* and *America* are binary variables that take a value equal to one when a language belongs to a certain area (and zero otherwise), and *Indoeuro*, *Afroasiatic*, *Nigercongo* and *Sinotibetan* are binary variables that take a value equal to one when a language belongs to a certain linguistic family.[9]

This way to introduce geographic and genetic factors in the relationship between phonemes per syllable versus syllables per word and words per clause is a relatively straightforward procedure to deal with the effects of those factors (which are here treated as “fixed effects”). It implies that the values of the relevant coefficients (which are now *c*(*10*) and *c*(*11*)) represent the effects of *Syll/Word* and *Word/Clause* on *Phon/Syll* that cannot be explained by the included geographic and genetic factors.

In the same way that eq. [16] tries to explain the different factors that influence the number of phonemes per syllable in our sample, syllables per word and words per clause can also be analyzed as dependent variables in regressions against the proposed binary variables. This implies writing equations like the following:

$$\begin{aligned} Syll/Word = {} & c(12)^{*}Europe + c(13)^{*}Africa + c(14)^{*}Westasia + c(15)^{*}Eastasia \\ & + c(16)^{*}America + c(17)^{*}Indoeuro + c(18)^{*}Afroasiatic \\ & + c(19)^{*}Nigercongo + c(20)^{*}Sinotibetan + c(21)^{*}Syll/Word \\ & + c(22)^{*}Word/Clause \end{aligned} \qquad [17];$$

$$\begin{aligned} Word/Clause = {} & c(23)^{*}Europe + c(24)^{*}Africa + c(25)^{*}Westasia \\ & + c(26)^{*}Eastasia + c(27)^{*}America + c(28)^{*}Indoeuro \\ & + c(29)^{*}Afroasiatic + c(30)^{*}Nigercongo \\ & + c(31)^{*}Sinotibetan + c(32)^{*}Syll/Word + c(33)^{*}Word/Clause \end{aligned} \qquad [18].$$

[9] These genetic variables have been included because they represent the four language families with a larger number of observations. The remaining families have only one or two observations within our 50-language sample, and due to that it is not convenient to create binary variables that represent them.

Due to the fact that the determinants of eqs [16]–[18] are basically the same, this is a case in which our analysis can be improved if we use "simultaneous-equation regressions". This method is relatively widespread in some social sciences such as economics, since it allows for procedures that single-equation regression analysis cannot deal with. The main one is the use of the correlations between the residuals of the three regression equations, through the so-called "seemingly unrelated regression" (SUR) procedure. This implies that, when estimating one equation, we also use information from the other equations, and that information can improve the precision and the statistical efficiency of the estimated coefficients.[10]

Equations [16]–[18] can therefore be run simultaneously, to see if we can find any statistical significance for the coefficients labeled as *c*(*10*), *c*(*11*), *c*(*21*), *c* (*22*), *c*(*32*) and *c*(*33*), which are the ones that measure the relationships between the different linguistic ratios. That analysis was performed using both ordinary least squares (OLS) and SUR.[11] Applying the same procedure described in Section 5.1, we used its results to calculate new partial correlation coefficients, which are the ones reported on Table 5.

Table 5: Partial correlation coefficients from simultaneous-equation regressions.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
OLS Regression			
Phonemes per syllable	1.0000	−0.3320	−0.1761
Syllables per word		1.0000	−0.6330
Words per clause			1.0000
SUR Regression			
Phonemes per syllable	1.0000	−0.5852	−0.4163
Syllables per word		1.0000	−0.8990
Words per clause			1.0000

[10] This procedure was originally proposed by Zellner (1962). In Coloma (2014), it has been used to detect possible correlations between phonological variables (consonant inventory size, vowel inventory size, tone distinctiveness and stress distinctiveness).

[11] The main results of these regressions are in Appendix 2. All of them were run using the statistical program EViews 3.1, but they could be replicated using other software packages. For readers accustomed to R, a useful package to run simultaneous-equation regressions is "systemfit". See Henningsen and Hamman (2007).

Note that the coefficients obtained when we use SUR are in all cases greater than the ones that we find when we use OLS (and they are also greater than the coefficients reported on Tables 2 and 3). This may be seen as a signal that the true negative correlation between the different linguistic ratios is higher than the one obtained when we perform a less sophisticated analysis, basically because in that analysis we do not take into account the interdependence between the linguistic ratios and other variables from the context in which those ratios appear (e. g., geographic and genetic factors).

6 Concluding remarks

After performing the numerical exercises described in Section 5, we can now write a few concluding remarks to summarize our results. The main one is that, using the dataset that we have assembled with 50 different versions of "The North Wind and the Sun", we have found that the syllable/word ratio is negatively correlated with the word/clause ratio, and that relationship is strong enough to appear as statistically significant in all the analyses that we have used for our data (see Table 6).

Table 6: Summary of the estimated correlation coefficients.

Concept	Phon/Syll vs. Syll/Word	Syll/Word vs. Word/Clause	Phon/Syll vs. Word/Clause
Simple correlation			
Coefficient	−0.2420	−0.6785	−0.0522
Probability	(0.0905)	(0.0000)	(0.7187)
Partial correlation			
Coefficient	−0.3781	−0.7132	−0.3036
Probability	(0.0074)	(0.0000)	(0.0340)
OLS with additional variables			
Coefficient	−0.3320	−0.6330	−0.1761
Probability	(0.0340)	(0.0000)	(0.2708)
SUR with additional variables			
Coefficient	−0.5852	−0.8990	−0.4163
Probability	(0.0001)	(0.0000)	(0.0068)

The correlation coefficient between phonemes per syllable and syllables per word, conversely, fails to be significant at a 5 % probability level when we compute it using the standard Pearson correlation technique. When calculated as a partial correlation coefficient, however, its absolute value increases, and it becomes statistically significant. This conclusion is maintained when we use simultaneous equations' methods in which we add binary variables related to geographic and genetic factors.

A similar phenomenon occurs with the relationship between the phoneme/syllable ratio and the word/clause ratio, as we can see on the last column of Table 6. In our dataset, these variables initially generate a small and insignificant product–moment correlation coefficient, but this changes considerably when we calculate partial correlation coefficients. Indeed, allowing for the interplay between the three linguistic ratios is enough to produce a coefficient that is statistically significant at a 5 % probability level, and this significance becomes stronger when we estimate that coefficient using seemingly unrelated regressions.

Note that, if we had only used standard product–moment coefficients to analyze the possible existence of negative correlation between our three linguistic ratios, we would have ended up with the conclusion that the only relevant relationship was the one between syllables per word and words per clause. Introducing a relatively simple method to deal with the interaction between the three ratios, and adding a set of binary geographic and genetic variables, we can see that the other negative correlations are also important, and this is especially true for the correlation between the phoneme/syllable and syllable/word ratios.

Our results can also be interpreted under the logic of a synergetic linguistics' model like the one that we propose in Section 4. In it, we assume that the three analyzed ratios are signals of different aspects of language complexity, and that their values are determined by an optimization process in which higher complexity is good to satisfy coding requirements but bad to satisfy economy requirements (related to production and memory effort). Therefore, when the external conditions in which a language develops (e. g., geographic and genetic factors) induce a certain linguistic measure to become larger, the postulated optimization process implies that this increase in complexity is somehow compensated by a reduction in the value of the other linguistic measures. These trade-offs become greater between linguistic ratios that are directly related (i. e., phoneme/syllable vs. syllable/word, and syllable/word vs. word/clause ratios), and smaller between ratios that are less related (i. e., phoneme/syllable vs. word/clause).

The conclusions obtained in this paper can be compared with results that appear in previous literature. The most compatible results are the ones gotten by Fenk-Oczlon and Fenk (1993, 2005), who also find negative and significant correlations between several linguistic ratios, and see them as evidences of complexity trade-offs. The likely cause of this coincidence is the fact that those authors use basically the same measures that we use in the present work, although their dataset is completely different from ours and many of the languages selected by them are also different.

The correlation between different cross-linguistic measures, conversely, appears to be insignificant in other studies such as Shosted (2006) and Nichols (2009). In those papers, the authors use theoretical measures of linguistic complexity instead of empirical measures based on text analysis. They also calculate correlations using standard Pearson coefficients only, and therefore do not include partial correlation coefficients. Those factors are probably the main causes for the discrepancies between their results and ours.

When some of the abovementioned factors are included in the analysis, some significant negative correlation is likely to appear between different cross-linguistic measures. That is the case of Wichmann et al. (2011), which finds a negative correlation between total inventory size and number of phonemes per word; Moran and Blasi (2014), which finds a negative correlation between vowel inventory size and number of phonemes per word; and Moscoso (2011), which finds a negative relationship between morphological structure and syntactic structure.

Acknowledgments: I thank Damián Blasi, Eric Campbell, Guiomar Ciapuscio, Bernard Comrie, John Du Bois, Matthew Gordon, Nick Lester, Fermín Moscoso, Franz Mueller, Verónica Nercesián, Frans Plank, Johan Rooryck, Patricia Schneider, Jorge Streb, Justin Watkins and two anonymous referees, for their useful comments. I especially thank Stefan Gries, editor of Corpus Linguistics and Linguistic Theory, and I also thank Federico Cápula, Helen Eaton, John Esling, Sameer Kahn and Kevin Schäfer, for their help in finding some of the sources used in this article. Finally, I thank participants in seminars held at CEMA University, the University of Buenos Aires, California State University at Fullerton, and the University of California at Santa Barbara (UCSB). Part of the research for this paper was conducted while I was working as a visiting scholar at UCSB.

Appendix 1: Linguistic ratios from “The North Wind and the Sun” dataset

Language	Family	Location	Phon/Syl	Syl/Word	Word/Clause
Amharic	Afro-Asiatic	Africa	2.5521	2.7553	11.75
Apache	Na-Dene	America	2.1287	2.3051	7.87
Arabic	Afro-Asiatic	West Asia	2.2488	2.5529	9.44
Arrernte	Pama-Nyungan	East Asia	2.2474	2.6575	6.08
Basque	Vasconic	Europe	2.1444	2.2530	11.86
Bemba	Niger-Congo	Africa	1.9333	2.8481	9.88
Bengali	Indo-European	West Asia	2.3299	1.8942	10.40
Burmese	Sino-Tibetan	East Asia	2.2901	3.1190	6.00
Cantonese	Sino-Tibetan	East Asia	2.7209	1.4176	9.10
Chickasaw	Muskogean	America	2.5761	3.2281	5.70
Dinka	Nilo-Saharan	Africa	1.9028	2.1022	13.70
English	Indo-European	Europe	2.6783	1.2655	12.56
French	Indo-European	Europe	2.1572	1.4722	12.00
Georgian	Kartvelian	West Asia	2.3616	2.5286	7.78
German	Indo-European	Europe	2.5111	1.6667	10.80
Greek	Indo-European	Europe	2.1577	2.1346	11.56
Hausa	Afro-Asiatic	Africa	2.2979	1.6988	13.83
Hebrew	Afro-Asiatic	West Asia	2.5288	2.3371	8.09
Hindi	Indo-European	West Asia	2.2452	1.6640	15.63
Hungarian	Uralic	Europe	2.2448	1.9200	10.00
Igbo	Niger-Congo	Africa	1.7115	1.9439	13.38
Irish	Indo-European	Europe	2.2809	1.3798	18.43
Japanese	Japonic	East Asia	1.9559	2.5506	9.89
Kabiye	Niger-Congo	Africa	1.9593	2.4286	10.11
Korean	Koreanic	East Asia	2.2952	2.7667	8.57
Malay	Austronesian	East Asia	2.3014	2.6795	9.75
Mandarin	Sino-Tibetan	East Asia	2.6815	1.6020	9.80
Mapudungun	Araucanian	America	2.3841	2.0133	8.33
Nara	Nilo-Saharan	Africa	2.3417	1.8426	9.82
Nepali	Indo-European	West Asia	2.1921	2.4105	10.56
Persian	Indo-European	West Asia	2.4897	2.1319	10.11
Portuguese	Indo-European	Europe	2.0430	1.8980	12.25
Quichua	Quechuan	America	2.1882	2.8830	8.55
Russian	Indo-European	Europe	2.5027	1.9278	10.78
Sahaptin	Penutian	America	2.4351	2.7018	7.13
Sandawe	Khoisan	Africa	2.1044	2.3038	8.78
Seri	Hokan	America	2.4504	1.5414	14.27
Shiwilu	Kawapanan	America	2.4121	3.2130	7.71

(*continued*)

(*continued*)

Language	Family	Location	Phon/Syl	Syl/Word	Word/Clause
Spanish	Indo-European	Europe	2.2021	1.9897	10.78
Tajik	Indo-European	West Asia	2.1810	2.4556	12.86
Tamil	Dravidian	West Asia	2.1468	3.1899	8.78
Tashlhiyt	Afro-Asiatic	Africa	2.5082	1.6053	8.44
Tausug	Austronesian	East Asia	2.4034	2.0877	9.50
Temne	Niger-Congo	Africa	2.1546	1.6560	11.36
Thai	Tai-Kadai	East Asia	2.7746	1.3206	11.91
Trique	Oto-Manguean	America	1.9617	1.7103	10.70
Turkish	Altaic	West Asia	2.3552	2.7727	7.33
Vietnamese	Austro-Asiatic	East Asia	2.8547	1.0000	16.71
Yine	Arawakan	America	2.3686	3.7460	6.30
Zapotec	Oto-Manguean	America	2.1234	1.7701	9.67

Appendix 2: Main results from simultaneous-equation regressions

Concept	OLS			SUR		
	Coefficient	*t*-statistic	*p*-value	Coefficient	*t*-statistic	*p*-value
Phoneme/Syllable equation						
Europe [c(1)]	2.750302	9.030173	0.0000	3.258497	12.57648	0.0000
Africa [c(2)]	2.694881	8.578103	0.0000	3.220800	12.04544	0.0000
Westasia [c(3)]	2.793696	8.953035	0.0000	3.327674	12.58630	0.0000
Eastasia [c(4)]	2.928329	9.760996	0.0000	3.442674	13.51247	0.0000
America [c(5)]	2.850239	9.677724	0.0000	3.363552	13.49440	0.0000
Indoeuro [c(6)]	0.052739	0.477201	0.6341	0.062481	0.641154	0.5227
Afroasiatic [c(7)]	0.221723	1.852404	0.0665	0.210593	1.992672	0.0486
Nigercongo [c(8)]	−0.205678	−1.433084	0.1545	−0.188250	−1.485452	0.1401
Sinotibetan [c(9)]	0.105424	0.729699	0.4670	0.041447	0.325537	0.7454
Syllable/Word [c(10)]	−0.154563	−2.197749	0.0299	−0.272465	−4.623149	0.0000
Word/Clause [c(11)]	−0.018470	−1.117133	0.2662	−0.043665	−3.066114	0.0027
R-squared	0.4260			0.3790		
Syllable/Word equation						
Europe [c(12)]	5.200970	6.557863	0.0000	6.999677	10.57539	0.0000
Africa [c(13)]	5.230044	6.680569	0.0000	7.058802	10.83656	0.0000
Westasia [c(14)]	5.500062	7.120577	0.0000	7.221758	11.19126	0.0000

(*continued*)

(*continued*)

Concept	**OLS**			**SUR**		
	Coefficient	***t*-statistic**	***p*-value**	**Coefficient**	***t*-statistic**	***p*-value**
Eastasia [c(15)]	5.341652	6.394532	0.0000	7.269296	10.45097	0.0000
America [c(16)]	5.382757	6.846725	0.0000	7.151802	10.90478	0.0000
Indoeuro [c(17)]	−0.062841	−0.264211	0.7921	0.153317	0.742032	0.4596
Afroasiatic [c(18)]	0.053080	0.198051	0.8433	0.205081	0.870139	0.3860
Nigercongo [c(19)]	−0.033348	−0.105460	0.9162	−0.137267	−0.492910	0.6230
Sinotibetan [c(20)]	−0.283525	−0.917274	0.3609	−0.319307	−1.177676	0.2413
Phoneme/Syllable [c(21)]	−0.712982	−2.197749	0.0299	−1.256853	−4.623149	0.0000
Word/Clause [c(22)]	−0.142613	−5.106862	0.0000	−0.202527	−9.719000	0.0000
R-squared	0.5847			0.5020		
Word/Clause equation						
Europe [c(23)]	19.76857	4.936155	0.0000	27.48218	8.199957	0.0000
Africa [c(24)]	20.51415	5.278330	0.0000	27.99005	8.603652	0.0000
Westasia [c(25)]	19.88516	4.837152	0.0000	28.14427	8.249618	0.0000
Eastasia [c(26)]	20.42860	4.888437	0.0000	28.47519	8.130431	0.0000
America [c(27)]	19.54471	4.737288	0.0000	27.78242	8.092900	0.0000
Indoeuro [c(28)]	1.514544	1.472593	0.1435	1.076125	1.195628	0.2343
Afroasiatic [c(29)]	0.278197	0.233891	0.8155	0.632122	0.602473	0.5480
Nigercongo [c(30)]	0.159021	0.113295	0.9100	−0.255523	−0.206318	0.8369
Sinotibetan [c(31)]	−2.074152	−1.540534	0.1261	−1.833498	−1.542564	0.1256
Phoneme/Syllable [c(32)]	−1.678762	−1.117133	0.2662	−3.968658	−3.066114	0.0027
Syllable/Word [c(33)]	−2.809959	−5.106862	0.0000	−3.990447	−9.719000	0.0000
R^2	0.5912			0.5336		

References

Altmann, Gabriel & Gejza Wimmer. 2007. Towards a unified derivation of some linguistic laws. In P. Grzybek (ed.), *Contributions to the science of text and language*, 329–338. Dordrecht: Springer.

Coloma, Germán. 2014. Towards a synergetic statistical model of language phonology. *Journal of Quantitative Linguistics* 21(2). 100–122.

Coloma, Germán. 2015. The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics* 28. 139–159.

Fenk-Oczlon, Gertraud & AugustFenk. 1985. The mean length of propositions is 7 plus minus 2 syllables, but the position of languages within this range is not accidental. In G. D'Ydevalle (ed.), *Cognition, information processing and motivation,* 355–359. Amsterdam: North Holland.

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

Fenk-Oczlon, Gertraud & August Fenk. 2005. Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In G. Fenk-Oczlon & C. Winkler (eds.), *Sprache und Natürlichkeit,* 75–86. Tübingen: Narr.

Henningsen, Arne & Jeff Hamann. 2007. Systemfit: A package for estimating systems of simultaneous equations in R. *Journal of Statistical Software* 23(4). 1–40.

IPA. 1999. *Handbook of the International Phonetic Association*. Cambridge: Cambridge University Press.

Köhler, Reinhard. 1987. System theoretical linguistics. *Theoretical Linguistics* 14(2–3). 241–257.

Köhler, Reinhard. 2005. Synergetic linguistics. In G. Altmann, R. Köhler & R. Piotrowski (eds.), *Quantitative linguistics: An international handbook*, 760–774. Berlin: De Gruyter.

Lowry, Richard. 2000. *Concepts and applications of inferential statistics*. Poughkeepsie: Vassar College Press.

Maddieson, Ian. 2007. Issues of phonological complexity: Statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. In M. Solé, P. Beddor & M. Ohala (eds.), *Experimental approaches to phonology*, 93–103. New York: Oxford University Press.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language complexity: Typology, contact and change,* 23–41. Amsterdam: John Benjamins.

Moran, Steven & Damián Blasi. 2014. Cross-linguistic comparison of complexity measures in phonological systems. In F. Newmayer & L. Preston (eds.), *Measuring grammatical complexity*, 217–240. New York: Oxford University Press.

Moscoso, Fermín. 2011. The universal "shape" of human language: Spectral analysis beyond speech. *Nature Precedings*, 6097-2.

Nettle, Daniel. 1995. Segmental inventory size, word length and communicative efficiency. *Linguistics* 33. 359–367.

Nichols, Johanna. 2009. Linguistic complexity: A comprehensive definition and survey. In G. Sampson, D. Gil & P. Trudgill (eds.), *Language complexity as an evolving variable*, 110–125. Oxford: Oxford University Press.

Oh, Yoon-Mi, François Pellegrino, Egidio Marsico& Christophe Coupé. 2013. A quantitative and typological approach to correlating linguistic complexity. In *Proceedings of the 5th Conference on Quantitative Investigations in Theoretical Linguistics*, 71–75. Leuven: University of Leuven.

Shosted, Ryan. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10. 1–40.

Vulanovic, Relja. 2007. On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics* 20. 399–427.

Wichmann, Soren, Taraka Rama & Eric Holman. 2011. Phonological diversity, word length and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15. 157–177.

Zellner, Arnold. 1962. An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *Journal of the American Statistical Association* 57. 348–368.

Data sources

Arvaniti, Amalia. 1999. Standard Modern Greek. *Journal of the International Phonetic Association* 29(2). 167–172.

Breen, Gavan & Veronica Dobson. 2005. Central Arrernte. *Journal of the International Phonetic Association* 35(2). 249–254.

Clynes, Adrian & David Deterding. 2011. Standard Malay (Brunei). *Journal of the International Phonetic Association* 41(2). 259–268.

Cruz-Ferreira, Madalena. 1999. Portuguese (European). In IPA (1999), 126–130.

Dawd, Abushush & Richard Hayward. 2002. Nara. *Journal of the International Phonetic Association* 32(2). 249–255.

DiCanio, Christian. 2010. Itunyoso Trique. *Journal of the International Phonetic Association* 40(2). 227–238.

Eaton, Helen. 2006. Sandawe. *Journal of the International Phonetic Association* 36(2). 235–242.

Fougeron, Cécile & Caroline Smith. 1999. French. In IPA (1999), 78–81.

Gordon, Matthew, Pamela Munro & Peter Ladefoged. 2001. Chickasaw. *Journal of the International Phonetic Association* 31(2). 287–290.

Hamann, Silke & Nancy Kula. 2015. Bemba. *Journal of the International Phonetic Association* 45 (1). 61–69.

Hargus, Sharon & Virginia Beavert. 2014. Northwest Sahaptin. *Journal of the International Phonetic Association* 44(3). 320–342.

Hayward, Katrina & Richard Hayward. 1999. Amharic. In IPA (1999), 45–50.

Hualde, José, Oihana Lujanbio & Juan Zubiri. 2010. Goizueta Basque. *Journal of the International Phonetic Association* 40(1). 113–127.

Ido, Shinji. 2014. Bukharan Tajik. *Journal of the International Phonetic Association* 44(1). 87–102.

Ikekeonwu, Clara. 1999. Igbo. In IPA (1999), 108–110.

Kahn, Sameer. 2010. Bengali (Bangladeshi Standard). *Journal of the International Phonetic Association* 40(2). 221–225.

Kanu, Sullay & Benjamin Tucker. 2010. Temne. *Journal of the International Phonetic Association* 40(2). 247–253.

Keane, Elinor. 2004. Tamil. *Journal of the International Phonetic Association* 34(1). 111–116.

Khatiwada, Rajesh. 2009. Nepali. *Journal of the International Phonetic Association* 39(3). 373–380.

Kirby, James. 2011. Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association* 41(3). 381–392.

Kohler, Klaus. 1999. German. In IPA (1999), 86–89.

Laufer, Asher. 1999. Hebrew. In IPA (1999), 96–99.

Lee, Hyun Bok. 1999. Korean. In IPA (1999), 120–123.

Lee, Wai-Sum & Eric Zee. 2003. Standard Chinese (Beijing). *Journal of the International Phonetic Association* 33(1). 109–112.

Majidi, Mohammad & Elmar Ternes. 1999. Persian (Farsi). In IPA (1999), 124–125.

Marlett, Stephen, Xavier Moreno & Genaro Herrera. 2005. Seri. *Journal of the International Phonetic Association* 35(1). 117–121.

Martínez, Eugenio, Ana Fernández & Josefina Carrera. 2003. Castilian Spanish. *Journal of the International Phonetic Association* 33(2). 255–260.

Masaquiza, Fanny & Stephen Marlett. 2008. Salasaca Quichua. *Journal of the International Phonetic Association* 38(2). 223–227.

Ní Chasaide, Ailbhe. 1999. Irish. In IPA (1999), 111–116.

Ohala, Manjari. 1999. Hindi. In IPA (1999), 100–103.

Okada, Hideo. 1999. Japanese. In IPA (1999), 117–119.

Padayodi, Cécile. 2008. Kabiye. *Journal of the International Phonetic Association* 38(2). 215–221.

Pickett, Velma, María Villalobos & Stephen Marlett. 2010. Isthmus (Juchitán) Zapotec. *Journal of the International Phonetic Association* 40(3). 365–372.

Remijsen, Bert & Caguor Manyang. 2009. Luanyjang Dinka. *Journal of the International Phonetic Association* 39(1). 123–124.

Roach, Peter. 2004. British English: Received Pronunciation. *Journal of the International Phonetic Association* 34(2). 239–245.

Ridouane, Rachid. 2014. Tashlhiyt Berber. *Journal of the International Phonetic Association* 44(2). 207–221.

Sadowsky, Scott, Héctor Painequeo, Gastón Salamanca & Heriberto Avelino. 2013. Mapudungun. *Journal of the International Phonetic Association* 43(1). 87–96.

Schuh, Russell & Lawan Yalwa. 1999. Hausa. In IPA (1999), 90–95.

Shosted, Ryan & Vakhtang Chikovani. 2006. Standard Georgian. *Journal of the International Phonetic Association* 36(2). 255–264.

Soderberg, Craig, Seymour Ashley & Kenneth Olson. 2012. Tausug (Suluk). *Journal of the International Phonetic Association* 42(3). 361–364.

Szende, Tamás. 1999. Hungarian. In IPA (1999), 104–107.

Thelwall, Robin &Akram Sa'adeddin. 1999. Arabic. In IPA (1999), 51–54.

Tingsabadh, Kalaya & Arthur Abramson. 1999. Thai. In IPA (1999), 147–150.

Tuttle, Siri & Merton Sandoval. 2002. Jicarilla Apache. *Journal of the International Phonetic Association* 32(1). 105–112.

Urquía, Rittma & Stephen Marlett. 2008. Yine. *Journal of the International Phonetic Association* 38(3). 365–369.

Valenzuela, Pilar & Carlos Gussenhoven. 2013. Shiwilu (Jebero). *Journal of the International Phonetic Association* 43(1). 97–106.

Watkins, Justin. 2001. Burmese. *Journal of the International Phonetic Association* 31(2). 291–295.

Yanushevskaya, Irena & Daniel Buncic. 2015. Russian. *Journal of the International Phonetic Association* 45(2). 221–228.

Zee, Eric. 1999. Cantonese. In IPA (1999), 58–60.

Zimmer, Karl & Orhan Orgun. 1999. Turkish. In IPA (1999), 154–156.

Bionote

Germán Coloma

Germán Coloma is a full-time professor at CEMA University, Buenos Aires, Argentina. Although his main area of research has to do with the use of statistical methods in economic analysis, he has published several papers about quantitative linguistics, especially in areas related to sociolinguistics, phonetics and linguistic typology. He has also been a visiting scholar at the UCSB Linguistics Department and holds a Ph.D. in economics from UCLA.