Studies in Linguistics and Literature
ISSN 2573-6434 (Print) ISSN 2573-6426 (Online)
Vol. 6, No. 4, 2022
www.scholink.org/ojs/index.php/sll

Original Paper

Correlation between Linguistic Measures: An Extended Analysis

Germán Coloma[1*]

[1] CEMA University, Buenos Aires, Argentina

[*] Germán Coloma, CEMA University, Buenos Aires, Argentina

Received: October 10, 2022 Accepted: October 20, 2022 Online Published: November 1, 2022

doi:10.22158/sll.v6n4p109 URL: http://dx.doi.org/10.22158/sll.v6n4p109

Abstract

In this paper we reconsider the results that appeared originally in Coloma (2017) about the possible existence of negative correlation between linguistic measures, using a newly-assembled database of 80 languages for which we have the same text (which is the fable known as "The North Wind and the Sun"). Most conclusions of the original paper become reinforced, especially the ones related to the existence of language complexity trade-offs. This is particularly clear when we look at partial correlation coefficients between three linguistic ratios (phonemes per syllable, syllables per word, and words per clause), especially when we use simultaneous-equation regression methods and instrumental variables.

Keywords

complexity trade-off, partial correlation, linguistic ratios, simultaneous-equation regression, instrumental variables

1. Introduction

In Coloma (2017), there is an analysis about the possible existence of negative correlation between linguistic measures, using the text of a relatively famous fable ("The North Wind and the Sun") translated into 50 different languages (Note 1). With those translations, a database was built, with information concerning several empirical measures for the text under analysis (phonemes per syllable, syllables per word, words per clause). That database also included other variables related to the typological characteristics of the languages, and to some "non-linguistic" factors (e.g., location of the languages, phylogenetic relationships, number of speakers).

The main conclusion of the abovementioned paper is that negative correlations exist and are significant in the context under analysis. They also seem to be partially hidden, because of possible interactions among different variables. As a consequence of that, it holds that the correlation coefficients seem to be higher and more significant if we take into account those interactions. In order to do that, a

combination of different alternative strategies was used, and that implied using partial correlation coefficients, simultaneous-regression equations, non-linguistic variables and typological variables. One limitation of the analysis, however, has to do with the database itself, which only has 50 observations. That limitation was due to the fact that, when calculations were performed, there were relatively few examples for the text that was used to compare the different languages, and many of those examples were about languages that were not different enough (in terms of their phylogenetic and/or geographic variation).

As several years have passed, we have been able to build another alternative database with 80 languages for which we have the text of "The North Wind and the Sun". The source of those languages is essentially the same one used for the original sample, i.e., it is the collection of "Illustrations of the IPA" published in IPA (1999) and in the *Journal of the International Phonetic Association* (Note 2). As that collection is now considerably larger, this new database has the advantage that it is more diverse, in the sense that it contains languages from more families (and not so many Indo-European languages).

In this paper, we use our newly-assembled database to perform essentially the same analyses that appear in Coloma (2017). First, we describe the basic characteristics of the database in terms of its scope of languages and the value of the calculated linguistic measures (section 2). Then, in section 3, we use those measures to compute correlation coefficients, using alternative methodologies. In section 4 we include additional variables in the analysis, to deal with geographic, phylogenetic and population factors, while in section 5 we introduce a procedure that replaces our linguistic measures by "instrumental variables". Finally, in section 6, we compare our results with the original ones, and state a few concluding remarks.

2. 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 (Note 3). It is therefore a unique case of a short text for which specialists in the phonetics of many different languages have identified the sounds, the phonemes, the syllables and the words of those languages.

In Coloma (2017), there is a database that relies on the text of "The North Wind and the Sun" translated into the following languages: Sahaptin, Eastern Apache, Chickasaw, Seri, Trique, Isthmus Zapotec, Ecuadorian Quichua, Shiwilu, Yine and Mapudungun (which are original of the American continent); Portuguese, Spanish, Basque, French, Irish, English, German, Russian, Hungarian and Greek (from Europe); Tashlhiyt Berber, Temne, Kabiye, Igbo, Hausa, Dinka, Nara, Amharic, Sandawe and Bemba (from Africa); Georgian, Turkish, Hebrew, Arabic, Persian, Tajik, Nepali, Hindi, Bengali and Tamil (from West Asia); and Japanese, Korean, Mandarin, Cantonese, Burmese, Thai, Vietnamese, Malay, Tausug and Arrernte (from East Asia and Australasia).

For this paper, we have built a new sample with 80 languages (see Figure 1). Thirty-six of them already appeared in the original database, and those are Amharic, Apache, Arabic, Arrernte, Basque, Bengali, Berber, Burmese, Chickasaw, Dinka, Georgian, Greek, Hausa, Hungarian, Igbo, Irish, Japanese, Kabiye, Korean, Malay, Mandarin, Mapudungun, Nara, Persian, Quichua, Sahaptin, Sandawe, Seri, Tamil, Tausug, Temne, Thai, Trique, Turkish, Vietnamese and Zapotec. Additionally, the new forty-four languages included are Abkhaz, Aingae, Albanian, Ambel, Bai, Basaa, Ende, Estonian, Fataluku, Hmong, Huehuetla Tepehua, Italian, Kalabari Ijo, Kazakh, Kera, Khongso, Kumiai, Kunama, Lizu, Mah Meri, Makasar, Malagasy, Mambay, Mojeño, Mono, Nen, Paiute, Panjabi, Pitjiantjatjara, Qanjobal, Qaqet, Seenku, Setswana, Shawi, Shipibo, Sumi, Swedish, Tamambo, Telugu, Tepehuan, Tera, Tongan, Ukrainian and Urarina.

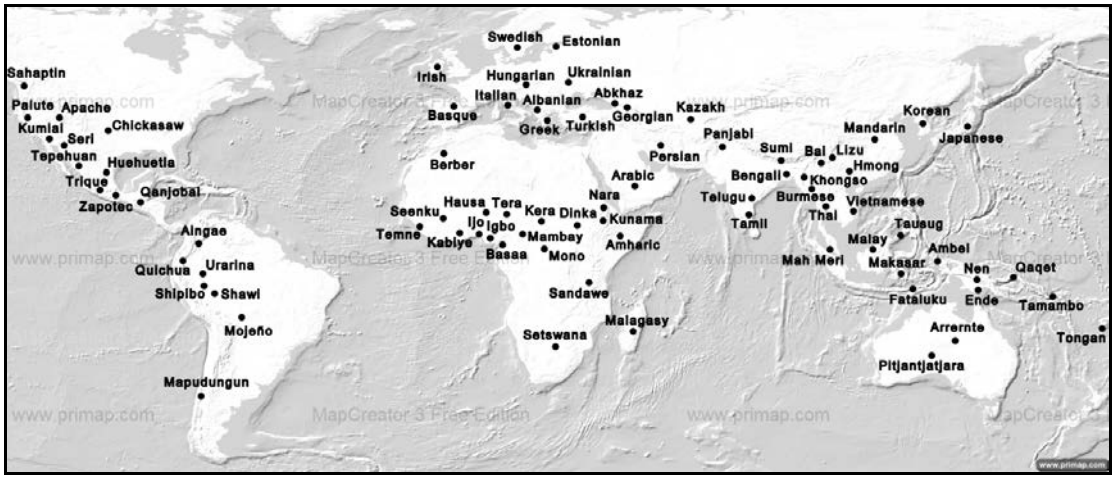

Figure 1. Location of the Languages Included in the Sample

One obvious difference between the original database and the newly-assembled one is that the latter has thirty more languages. Besides, its composition is clearly more diverse, since it includes examples from 40 different language families. In contrast, the original sample had languages belonging to only 26 families, and 13 of those languages (26%) were Indo-European. Moreover, some pairs of languages were very similar between them (as was the case of Spanish and Portuguese, Persian and Tajik, Mandarin and Cantonese, etc.). In our newly-assembled database, conversely, all languages belong to different “subfamilies” (Note 4).

The basic statistics computed for these samples of languages come from counting the number of phonemes, syllables, words and clauses included in the translation of “The North Wind and the Sun” for each of those languages. With those figures, we can calculate a series of linguistic ratios, which basically are the phoneme/syllable ratio, the syllable/word ratio and the word/clause ratio. These ratios can be seen as measures of different aspects of language complexity (Note 5). For example, the phoneme/syllable ratio may be linked to the degree of phonological complexity at the level of the

syllable, while the number of syllables per word can be associated to (morphological) complexity at the level of the word. Finally, the word/clause ratio can be seen as an empirical measure of syntactic complexity, since clauses with more words tend to be related to situations in which syntax is more complex.

In the sample that we use in this paper, the phoneme/syllable ratio goes from a minimum of 1.7115 (for the case of Igbo, a Niger-Congo language spoken in Nigeria) to a maximum of 2.9024 (for the case of Kumiai, a Yuman language spoken in the Mexican/US border), in a context in which the average number of phonemes per syllable is 2.2491. The syllable/word ratio, conversely, goes from a minimum of 1 (for the Vietnamese language) to a maximum of 3.6 (for Telugu, a Dravidian language spoken in India), in a context where the average number of syllables per word is 2.1541. Finally, the minimum word/clause ratio is equal to 4.5, and corresponds to Paiute (which is a Uto-Aztecan language spoken in the United States), while the maximum word/clause ratio in the sample is 21.67, and corresponds to the Tongan language (in a context in which the average number of words per clause is 11.15) (Note 6).

3. Standard and Partial Correlation Coefficients

The linguistic measures described in the previous section can be correlated between themselves. As we have three ratios (phonemes per syllable, syllables per word, and words per clause), we can find three correlation coefficients, which are the ones that appear on Table 1.

Table 1. Standard Correlation Coefficients

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000		
Syllables per word	-0.2384	1.0000	
Words per clause	-0.1004	-0.5919	1.0000

The basic meaning of these correlation coefficients is that each linguistic ratio is negatively related with the other two ratios. This gives a hint of possible trade-offs between those language complexity measures, in the sense that, on average, a language that is more complex in a certain dimension tends to be simpler in another dimension. For example, the text of “The North Wind and the Sun” translated into Hmong (Hmong-Mien, spoken in China) has an average of 1.0637 syllables per word, and an average of 15.7 words per clause. Conversely, the same text in Chickasaw (Muskogean, spoken in the United States) has an average of 3.2281 syllables per word but only 5.7 words per clause. In this case, this trade-off seems to create an actual trend for the whole database, as can be seen in Figure 2 (where each language is represented by a point in a space that maps syllables per word versus words per clause).

The absolute values of the correlation coefficients reported on Table 1 are also related to their statistical

significance. For instance, correlation between syllables per word and words per clause (r = -0.5919) is significantly different from zero at a 1% probability level, since its probability value (p = 0.0000) is smaller than 1%. On the contrary, the correlation coefficient between phonemes per syllable and syllables per word (r = -0.2384) is only significant at a 5% level, since its probability value (p = 0.0332) is smaller than 5% but larger than 1%. Finally, the correlation coefficient between phonemes per syllable and words per clause (r = -0.1004) fails to be significant at any reasonable probability level, since its probability value (p = 0.3754) is above 10%.

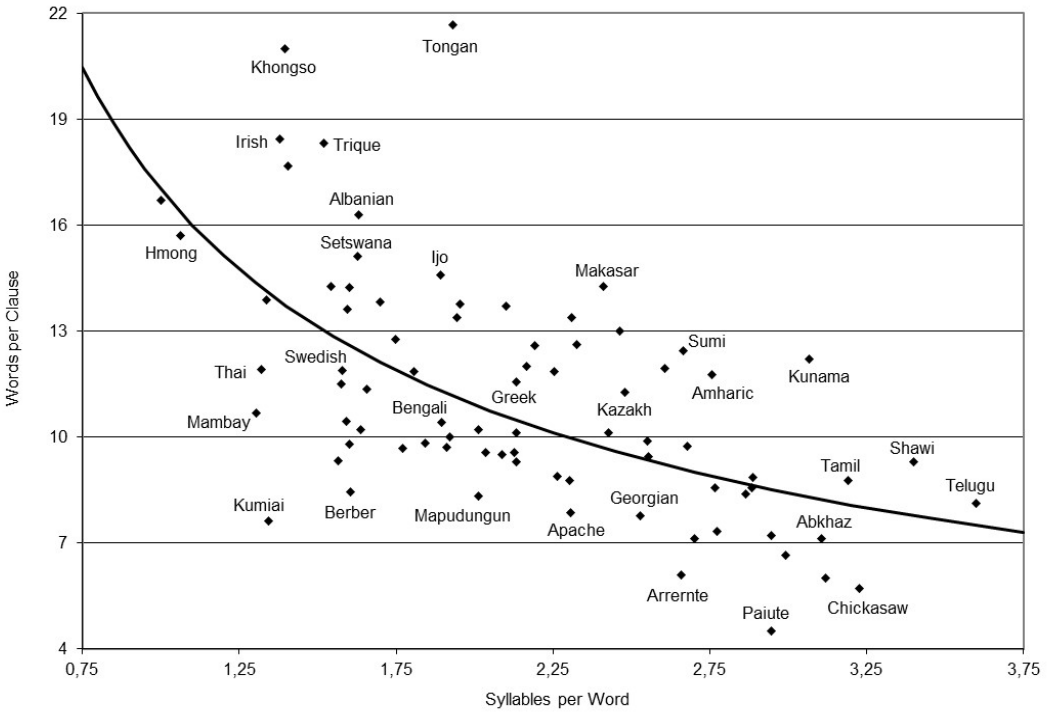

Figure 2. Relationship between Syllables per Word and Words per Clause

In Coloma (2017), there is an interesting empirical result related to correlation between complexity measures, which appears by comparing the standard or "product-moment" correlation coefficients with their corresponding "partial correlation coefficients". The standard correlation coefficients (which are the ones reported on Table 1) are calculated using information of the variables that one wishes to correlate, but they do not use any information about additional variables that may have influence on the magnitudes that are compared. Conversely, the partial correlation coefficients are calculated "controlling for" (i.e., using information about) other variables that may be themselves correlated with the two variables that we wish to study.

To calculate a partial correlation coefficient, it is necessary to eliminate the possible effect of other factors on the two variables that we wish to correlate, using some statistical procedure. One possibility

is to begin with a complete correlation matrix for all the variables under analysis (which in our case are only three variables), and then invert that matrix. Once we do that, we can use the following formula:

$$r_{xy} = -\frac{i_{xy}}{\sqrt{i_{xx} \cdot i_{yy}}} \quad (1);$$

where i_{xy} is the coefficient that corresponds to the pair of variables *x* and *y* in the inverse correlation matrix, and i_{xx} and i_{yy} are the coefficients that correspond to those variables in the main diagonal of the same inverse correlation matrix (Note 7). This process of matrix inversion is actually one of the possibilities that can be used to obtain partial correlation coefficients. Another one is to run three regression equations, in which each variable is regressed against a constant and the other two variables. Both procedures have the same goal, which is pulling out the effects that the remaining variable may have on each pair of variables that we are interested in.

If we apply the regression procedure in this case, we need to run three regression equations that consist of the following linear functions:

*Phoneme/Syllable = c(1) +c(2)*Syllable/Word +c(3)*Word/Clause* (2) ;

*Syllable/Word = c(4) +c(5)*Phoneme/Syllable +c(6)*Word/Clause* (3) ;

*Word/Clause = c(7) +c(8)*Phoneme/Syllable +c(9)*Syllable/Word* (4) ;

where *Phoneme/Syllable*, *Syllable/Word* and *Word/Clause* are our three linguistic ratios, and *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.

Table 2. OLS Regression Results to Calculate Partial Correlation Coefficients

Concept	Coefficient	t-Statistic	Probability
Phoneme/Syllable Equation			
Constant [c(1)]	2.93436	15.19583	0.0000
Syllable/Word [c(2)]	-0.18285	-3.51006	0.0005
Word/Clause [c(3)]	-0.02614	-2.84651	0.0048
R-squared	0.1466		
Syllable/Word Equation			
Constant [c(4)]	5.07345	9.57201	0.0000
Phoneme/Syllable [c(5)]	-0.75439	-3.51006	0.0005
Word/Clause [c(6)]	-0.10969	-7.25756	0.0000
R-squared	0.4399		
Word/Clause Equation			
Constant [c(7)]	27.31484	8.20534	0.0000
Phoneme/Syllable [c(8)]	-3.64179	-2.84651	0.0048
Syllable/Word [c(9)]	-3.70326	-7.25756	0.0000
R-squared	0.4122		

When we run our regression equations using Ordinary Least Squares (OLS) (Note 8), we obtain the results that appear on Table 2. With those results, the partial correlations between the different linguistic ratios can be calculated by using the following formula:

$$r_{xy} = -\sqrt{c_{xy} \cdot c_{yx}} \quad (5);$$

where r_{xy} is the partial correlation coefficient between variable x and variable y, c_{xy} is the regression coefficient of variable y in variable x's equation, and c_{yx} is the regression coefficient of variable x in variable y's equation. Note that in this formula we assume that, as both regression coefficients are negative, the corresponding partial correlation coefficient must also be negative.

Applying our formula to the results reported on Table 2, it is possible to obtain the partial correlation coefficients that are shown on Table 3. If we compare those results with the ones that appear on Table 1, we see that the three partial correlation coefficients are higher than their corresponding standard correlation coefficients (Note 9). This is also linked to a larger statistical significance, which in this case is given by the fact that the three calculated coefficients are significant at a 1% probability level ("p = 0.0007"; "p = 0.0054" and "p = 0.0000").

Table 3. Partial Correlation Coefficients Using OLS

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000		
Syllables per word	-0.3714	1.0000	
Words per clause	-0.3086	-0.6373	1.0000

The idea of calculating partial correlation coefficients using a system of linear equations can be complemented with the use of a methodology known as "simultaneous-equation regression". This consists of running the three regression equations at the same time, making use of some relationships that can be found between the different equations. One of those relationships is the covariance between the results of the different equations, in particular the one between the "residuals" of the equations (i.e., between the part of the variation of each dependent variable that cannot be explained by the regression procedure).

If we introduce that element in the estimation of the simultaneous-equation regressions under analysis, then we will be running a system of "Seemingly Unrelated Regressions" (SUR). This is a procedure which is relatively widespread in some social sciences such as economics, and it implies that, when we are estimating each equation, we also use information from the other equations. That information can improve the precision and the statistical efficiency of the estimated coefficients (Note 10).

Table 4. Partial Correlation Coefficients Using SUR

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000		
Syllables per word	-0.6749	1.0000	
Words per clause	-0.6143	-0.9141	1.0000

Of course, the estimation of Equations 2, 3 and 4 using SUR produces a set of coefficients that can also be used to calculate partial correlations. In order to do that, we should use equation 5 with the newly-estimated regression coefficients. The new partial correlation coefficients are the ones reported on Table 4.

As we can see, the partial correlation coefficients obtained with this methodology are even larger than the ones reported on Table 3. All of them are also statistically significant at a 1% probability level (since their corresponding probability values are all equal to 0.0000). This may be interpreted as a signal that the true negative correlations between the different linguistic ratios are higher than the ones obtained when we use a simpler approach.

4. Geographic, Phylogenetic and Demographic Variables

In Coloma (2017, 2020), there is an extension of the analysis that explores the possibility of calculating partial correlation coefficients controlling for additional variables related to geographic, phylogenetic and demographic factors. This is performed by running a regression-equation system that includes those additional variables, such as the following one:

*Phoneme/Syllable = c(1)*Europe +c(2)*Africa +c(3)*Westasia +c(4)*Eastasia +c(5)*Australasia +c(6)*America +c(7)*Indoeuropean +c(8)*Afroasiatic +c(9)*Nigercongo +c(10)*Sinotibetan +c(11)*Austronesian +c(12)*Major +c(13)*Syllable/Word +c(14)*Word/Clause* (6);

*Syllable/Word = c(21)*Europe +c(22)*Africa +c(23)*Westasia +c(24)*Eastasia +c(25)*Australasia +c(26)*America +c(27)*Indoeuropean +c(28)*Afroasiatic +c(29)*Nigercongo +c(30)*Sinotibetan +c(31)*Austronesian +c(32)*Major +c(33)*Phoneme/Syllable +c(34)*Word/Clause* (7);

*Word/Clause = c(41)*Europe +c(42)*Africa +c(43)*Westasia +c(44)*Eastasia +c(45)*Australasia +c(46)*America +c(47)*Indoeuropean +c(48)*Afroasiatic +c(49)*Nigercongo +c(50)*Sinotibetan +c(51)*Austronesian +c(52)*Major +c(53)*Phoneme/Syllable +c(54)*Syllable/Word* (8);

where *Europe*, *Africa*, *Westasia*, *Eastasia*, *Australasia* and *America* are binary variables that take a value equal to one when a language belongs to a certain region (and zero otherwise); *Indoeuropean*, *Afroasiatic*, *Nigercongo*, *Sinotibetan* and *Austronesian* are variables that take a value equal to one when a language belongs to a certain linguistic family (Note 11); and *Major* is a variable that takes a value equal to one when a language is spoken by more than 5 million people (Note 12).

As we see, this new set of equations is basically equivalent to the one formed by Equations 2, 3 and 4, with the addition of twelve binary variables that are useful to include some factors that may influence language complexity but are essentially "non-linguistic". As this set of equations is also a system of simultaneous regressions, the seemingly-unrelated regression procedure mentioned in the previous section can also be applied here. The result is a set of 42 regression coefficients, which are the ones that appear on Table 5.

Table 5. Regression Results with Additional Variables Using SUR

Variable / Concept	Phoneme/Syllable Eq.		Syllable/Word Eq.		Word/Clause Eq.	
	Coefficient	**Probab.**	**Coefficient**	**Probab.**	**Coefficient**	**Probab.**
Europe	3.57281	0.0000	6.80414	0.0000	40.90643	0.0000
Africa	3.47449	0.0000	6.68302	0.0000	40.12009	0.0000
Westasia	3.58349	0.0000	6.89715	0.0000	40.91633	0.0000
Eastasia	3.55278	0.0000	6.54781	0.0000	39.93493	0.0000
Australasia	3.52532	0.0000	6.71792	0.0000	39.71263	0.0000
America	3.59446	0.0000	6.86332	0.0000	40.61706	0.0000
Indoeuropean	-0.09450	0.3510	-0.31181	0.1036	-0.56068	0.6747
Afroasiatic	0.04449	0.6748	-0.11398	0.5821	-0.28681	0.8397
Nigercongo	-0.26299	0.0130	-0.53837	0.0082	-2.10957	0.1392
Sinotibetan	-0.04579	0.6532	0.08098	0.6863	0.61831	0.6499
Austronesian	-0.02912	0.7731	0.01634	0.9341	1.34720	0.3079
Major	0.12036	0.0609	0.26386	0.0324	0.91563	0.2859
Phoneme/Syllable			-1.44447	0.0000	-7.75553	0.0000
Syllable/Word	-0.37431	0.0000			-5.52662	0.0000
Word/Clause	-0.04381	0.0000	-0.12047	0.0000		
R-squared	0.2835		0.5305		0.3613	

If we now use Equation 5 with the new regression coefficients (i.e., with the ones that correspond to the linguistic ratios in each of the three equations), we can obtain new partial correlation coefficients, which are the ones reported on Table 6. Once again, these coefficients are statistically significant at a 1% probability level, and this can be seen as a signal that the negative correlation phenomena that we found are still important when we control for the interaction between linguistic ratios and different geographic, phylogenetic and demographic factors.

Table 6. Partial Correlation Coefficients with Additional Variables Using SUR

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000		
Syllables per word	-0.7353	1.0000	
Words per clause	-0.5829	-0.8159	1.0000

5. Use of Instrumental Variables

An additional sophistication that can be included in our analysis is the use of the so-called "instrumental variables". These are helpful when we have a situation of endogeneity between the variables that we wish to correlate. In this case, for example, we are running a system of regressions whose first equation has *Phoneme/Syllable* as the dependent variable, while *Syllable/Word* is considered to be an independent variable. But this situation is reversed in the second equation, since *Syllable/Word* is the dependent variable there, while *Phoneme/Syllable* is one of the independent variables.

This fact, and a similar one that occurs with the third equation of our system, is considered to be a violation of the statistical assumptions embedded in the logic of the least-square regression methods, which assume that the independent variables must not be influenced by other variables of the system (i.e., that they must be "exogenous"). If we are estimating a regression equation in which some dependent variables are actually endogenous, that situation has to be solved using new variables that replace the original ones. These are the instrumental variables that we have to create, and they must be variables that have a relationship with the original ones but that, at the same time, are exogenous to the statistical problem under analysis (Note 13).

One relatively straightforward way to deal with the endogeneity problem in this case is to use typological variables that describe some characteristics of the languages, but are independent of the text under analysis. These can be variables related to the phonology of those languages (e.g., consonant inventory, vowel inventory, number of tones) or to their morphosyntax (e.g., number of cases, number of genders, number of verbal inflection categories). They can also be binary variables such as their level of syllable complexity (complex versus simple), use of phonological stress (distinctive versus non-distinctive), order of words (object-verb versus verb-object, adjective-noun versus noun-adjective), morphosyntactic alignment (accusative versus non-accusative) and degree of morphological synthesis (isolating versus concatenative).

In order to create all those variables for the 80 languages in our sample, we relied on the same sources of the text of "The North Wind and the Sun" (i.e., on the corresponding illustrations of the IPA) and also on information found in the World Atlas of Language Structures (Dryer & Haspelmath, 2013) (Note 14). Then we used those variables for a statistical procedure known as "Three-Stage Least Squares" (3SLS), which implies replacing the original endogenous variables (in our case,

Phoneme/Syllable, *Syllable/Word* and *Word/Clause*) for linear combinations of exogenous variables (Note 15).

The way to obtain linear combinations to replace the original endogenous variables is to run a new set of regression equations, in which *Phoneme/Syllable*, *Syllable/Word* and *Word/Clause* are regressed against the twelve typological variables and the twelve geographic, phylogenetic and population variables described in the previous section. Once we have the results of those regressions, the estimated coefficients (see Table 7) can be used to create new variables (that come from multiplying those coefficients by the values of the corresponding exogenous variables). With this we generate three instrumental variables (that we can call *Phonême/Syllâble*, *Syllâble/Wôrd* and *Wôrd/Claûse*) that replace the original variables of the regression systems but are at the same time completely exogenous to those systems.

Table 7. OLS Regression Results to Create Instrumental Variables

Variable / Concept	Phoneme/Syllable Eq.		Syllable/Word Eq.		Word/Clause Eq.	
	Coefficient	Probab.	Coefficient	Probab.	Coefficient	Probab.
Europe	2.23426	0.0000	1.88463	0.0000	12.73889	0.0001
Africa	2.36568	0.0000	2.08395	0.0000	9.81099	0.0019
Westasia	2.30895	0.0000	2.33080	0.0000	9.14239	0.0031
Eastasia	2.46568	0.0000	1.84920	0.0001	8.14629	0.0108
Australasia	2.41534	0.0000	2.04394	0.0000	9.36217	0.0002
America	2.39387	0.0000	2.43096	0.0000	7.72587	0.0010
Indoeuropean	-0.08083	0.5601	-0.51961	0.0451	2.22159	0.2268
Afroasiatic	0.04538	0.7536	-0.18799	0.4846	0.71410	0.7088
Nigercongo	-0.26708	0.0548	-0.30270	0.2401	0.87857	0.6312
Sinotibetan	-0.12161	0.3460	0.25798	0.2825	0.49828	0.7701
Austronesian	-0.12960	0.3885	-0.01960	0.9440	4.00995	0.0447
Major	0.08039	0.3271	0.26443	0.0839	-0.87142	0.4217
Consonants	-0.00069	0.8713	0.00454	0.5658	0.01542	0.7838
Vowels	0.00691	0.3901	0.00019	0.9897	0.05758	0.5879
Tones	0.00251	0.9246	-0.16023	0.0014	1.17220	0.0010
Stress	0.00004	0.9996	0.32539	0.0295	-0.30067	0.7758
Complex Syllables	0.20762	0.0084	-0.28071	0.0540	-0.32804	0.7503
Cases	-0.00550	0.7174	0.04951	0.0808	-0.36129	0.0733
Genders	-0.00483	0.8494	0.03488	0.4608	-0.17966	0.5932
Inflections	-0.02913	0.0694	-0.00873	0.7685	0.06171	0.7701
Object-Verb	-0.05725	0.4797	0.11572	0.4424	0.43932	0.6817

Adjective-Noun	0.00507	0.9409	-0.10221	0.4226	1.12597	0.2149
Accusative	-0.01515	0.8331	0.04448	0.7394	-1.24974	0.1900
Isolating	-0.03341	0.7205	0.08030	0.6438	-1.49095	0.2284
R-squared	0.3864		0.6626		0.4691	

The next step is running the actual set of equations that we are interested in (for example, the one formed by Equations 6, 7 and 8) but using the instrumental variables *Phonême/Syllâble*, *Syllâble/Wôrd* and *Wôrd/Claûse* instead of the original ones. If we introduce a seemingly unrelated regression procedure in this last stage, what we obtain is a three-stage least-square estimation whose coefficients can in turn be used to calculate new partial correlation coefficients. Those correlation coefficients are the ones shown on Table 8.

Table 8. Partial Correlation Coefficients with Additional Variables Using 3SLS

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1.0000		
Syllables per word	-0.9455	1.0000	
Words per clause	-0.8945	-0.9873	1.0000

As we did in the previous sections, we can now look at the values of the different correlation coefficients and their statistical significance. We see that they are all higher than the ones obtained when we used ordinary least squares and SUR, and all of them are significant at a 1% probability level. This can be considered as a signal that the negative correlation coefficients that were found before are robust to the removal of possible endogeneity biases that the data may have.

6. Conclusions and Comparison with Previous Results

The results reported in the previous sections, obtained using a database of 80 languages for which we have the text of “The North Wind and the Sun”, can be compared with the results that appear in Coloma (2017) for the original database of 50 languages. Performing that comparison (see Table 9), we can conclude that several stylized facts remain the same. For example, for both samples it holds that the partial correlation coefficients are higher than their corresponding standard correlation coefficients, and those coefficients tend to increase even more when we use an estimation method based on Seemingly Unrelated Regressions (SUR) or Three-Stage Least Squares (3SLS).

Table 9 also shows that the “ranking” of the correlation coefficients is unaltered by the use of different statistical methodologies. The highest coefficient is always the one that relates *Syllable/Word* with *Word/Clause*, followed by the coefficient that relates *Phoneme/Syllable* with *Syllable/Word*, while the correlation coefficient between *Phoneme/Syllable* and *Word/Clause* is always the one with the lowest

absolute value.

Concerning the use of instrumental variables, the results are not exactly the same. For the old database used in Coloma (2017), the correlation coefficient between phonemes per syllable and words per clause (r = -0.2350) is not statistically significant if calculated using a 3SLS methodology (Note 16). This is not the case with the newly-assembled database, for which that coefficient is "r = -0.8945" and is actually significant at any reasonable probability level. This result is important because 3SLS coefficients have the property that they are consistent and unbiased, and are therefore less likely to be influenced by stochastic shocks. As the new database is larger and more diverse than the original one, the new figures are probably more representative than the old ones, and they show that the implicit complexity trade-offs that are behind the negative correlation coefficients are actually more robust than the ones obtained using a smaller sample.

Table 9. Comparison of Results

Concept	Old database		New database	
	Coefficient	**Probab.**	**Coefficient**	**Probab.**
Standard correlation				
Phoneme/Syllable vs. Syllable/Word	-0.2420	0.0905	-0.2384	0.0332
Phoneme/Syllable vs. Word/Clause	-0.0522	0.7187	-0.1004	0.3754
Syllable/Word vs. Word/Clause	-0.6785	0.0000	-0.5919	0.0000
Partial correlation with OLS				
Phoneme/Syllable vs. Syllable/Word	-0.3781	0.0074	-0.3714	0.0007
Phoneme/Syllable vs. Word/Clause	-0.3036	0.0340	-0.3086	0.0054
Syllable/Word vs. Word/Clause	-0.7132	0.0000	-0.6373	0.0000
Partial correlation with SUR				
Phoneme/Syllable vs. Syllable/Word	-0.6730	0.0000	-0.6749	0.0000
Phoneme/Syllable vs. Word/Clause	-0.6047	0.0000	-0.6143	0.0000
Syllable/Word vs. Word/Clause	-0.9486	0.0000	-0.9141	0.0000
Partial correlation with additional variables (SUR)				
Phoneme/Syllable vs. Syllable/Word	-0.5852	0.0001	-0.7353	0.0000
Phoneme/Syllable vs. Word/Clause	-0.4163	0.0068	-0.5829	0.0000
Syllable/Word vs. Word/Clause	-0.8990	0.0000	-0.8159	0.0000
Partial correlation with additional variables (3SLS)				
Phoneme/Syllable vs. Syllable/Word	-0.5325	0.0003	-0.9455	0.0000
Phoneme/Syllable vs. Word/Clause	-0.2350	0.1391	-0.8945	0.0000
Syllable/Word vs. Word/Clause	-0.9361	0.0000	-0.9873	0.0000

After performing different kinds of calculations and estimations, and comparing them with the ones obtained for the original database, it is also possible to derive some additional conclusions and comments. The more prominent conclusion is that the language complexity trade-offs that were detected in the original study also appear in this paper. The main differences between the original results and the newly-obtained ones, however, are the following:

a) The correlation coefficients between phonemes per syllable and words per clause are all larger in the new database than in the old one.

b) The correlation coefficients between phonemes per syllable and syllables per word are also larger in the new database when we control for additional non-linguistic variables. When we do not control for those variables, the obtained coefficients are almost identical when we use the old and the new database.

c) Conversely, the correlation coefficients between syllables per word and words per clause are generally smaller in the new database than in the old one, except for the case in which we use instrumental variables to remove possible endogeneity biases.

Acknowledgement

I thank Christian Bentz, Stefan Gries, Jan Macutek, Alison Tompkins and one anonymous reviewer, for their comments on previous versions of this paper. I also thank Oraimar Socorro, for her help in finding many of the papers that were used as data sources for this article. All remaining errors are mine.

References

Baltagi, B. (2011). *Econometrics* (5th ed.). Berlin: Springer.

Bentz, C., Gutiérrez, X., Sozinova, O., & Samardzic, T. (2022). Complexity Trade-Offs and Equi-Complexity in Natural Languages: A Meta-Analysis. *Linguistics Vanguard*, in press.

Coloma, G. (2014). Towards a Synergetic Statistical Model of Language Phonology. *Journal of Quantitative Linguistics*, *21*, 100-122.

Coloma, G. (2016). *A Simultaneous-Equation Regression Model of Language Complexity Trade-Offs*. Buenos Aires: CEMA University Working Paper No. 597.

Coloma, G. (2017). The Existence of Negative Correlation between Linguistic Measures across Languages. *Corpus Linguistics and Linguistic Theory*, *13*, 1-26.

Coloma, G. (2020). *Language Complexity Trade-Offs Revisited*. Buenos Aires: CEMA University Working Paper No. 721.

Dryer, M., & Haspelmath, M. (2013). *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Fenk-Oczlon, G., & Fenk, A. (2008). Complexity Trade-Offs Between the Subsystems of Language. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language Complexity: Typology, Contact and*

Change (pp. 43-65). Amsterdam: John Benjamins.

IPA. (1912). *Principles of the International Phonetic Association*. Paris: International Phonetic Association.

IPA. (1949). *Principles of the International Phonetic Association*. London: University College.

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

Lichtman, K., Chang, S., Kramer, J., Crespo, C., Hallett, J., Huensch, A., & Morales, A. (2010) IPA Illustration of Q'anjob'al. *Studies in the Linguistic Sciences*, University of Illinois Working Paper.

Rasinger, S. (2013). *Quantitative Research in Linguistics* (2nd ed.). London: Bloomsbury.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias. *Journal of the American Statistical Association*, *57*, 348-368.

Zellner, A., & Theil, H. (1962). Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations. *Econometrica*, *30*, 54-78.

Notes

Note 1. See also Coloma (2020).

Note 2. We have also included one example taken from a different source (Lichtman et al., 2010).

Note 3. See, for example, IPA (1912), IPA (1949) and IPA (1999).

Note 4. For example, the nine Indo-European languages included in the new sample are Albanian (Albanic), Bengali (East Indic), Greek (Hellenic), Irish (Celtic), Italian (Romance), Panjabi (Northwest Indic), Persian (Iranian), Swedish (Germanic) and Ukrainian (Slavic).

Note 5. This interpretation has a long tradition in the quantitative linguistics' literature. See, for example, Fenk-Oczlon & Fenk (2008) and Bentz et al. (2022).

Note 6. For the complete list of values of the linguistic ratios, see Appendix 1.

Note 7. For an alternative explanation of the concept or partial correlation, see Rasinger (2013), chapter 7.

Note 8 .These regressions, and all the other ones reported in this paper, were run using the statistical program EViews 9.

Note 9. This result is equivalent to the one reported in Coloma (2017).

Note 10. This procedure was originally proposed by Zellner (1962), and is used in Coloma (2014) and Coloma (2017). For an explanation of its statistical properties, see Baltagi (2011), chapter 10.

Note 11. These are actually the five families that have six or more languages in the sample. Eight additional families have two languages each, and those are the Austro-Asiatic, Dravidian, East Sudanic, Oto-Manguean, Pama-Nyungan, Turkic, Uralic and Uto-Aztecan families. See Appendix 1.

Note 12. Due to this definition, the "major languages" in our sample are Albanian, Amharic, Arabic, Bengali, Burmese, Georgian, Greek, Hausa, Hungarian, Igbo, Italian, Japanese, Kazakh, Korean,

Malagasy, Malay, Mandarin, Panjabi, Persian, Setswana, Swedish, Tamil, Telugu, Thai, Turkish, Ukrainian and Vietnamese.

Note 13. For a more complete explanation of this, see Baltagi (2011), chapter 11.

Note 14. For the complete list of the values of the typological exogenous variables, see Appendix 2 and Appendix 3.

Note 15. This procedure was originally proposed by Zellner and Theil (1962). It is also used in Coloma (2016).

Note 16. This result is not reported in the original paper, that did not include an analysis of the endogeneity problem using instrumental variables.

Appendixes

Appendix 1. Linguistic Ratios

Language	Family	Region	Phon/Syll	Syll/Word	Word/Cl
Abkhaz	NW Caucasian	West Asia	2,1469	3,1053	7,13
Aingae	Cofan	America	2,0260	2,8657	8,38
Albanian	Indo-European	Europe	2,1613	1,6316	16,29
Ambel	Austronesian	Australasia	2,1162	1,9127	9,69
Amharic	Afro-Asiatic	Africa	2,5521	2,7553	11,75
Apache (Eastern)	Na-Dené	America	2,1287	2,3051	7,87
Arabic	Afro-Asiatic	West Asia	2,2488	2,5529	9,44
Arrernte	Pama-Nyungan	Australasia	2,2474	2,6575	6,08
Bai	Sino-Tibetan	East Asia	2,0546	1,5913	10,45
Basaa	Niger-Congo	Africa	2,2752	1,4057	17,67
Basque	Vasconic	Europe	2,1444	2,2530	11,86
Bengali	Indo-European	West Asia	2,3299	1,8942	10,40
Berber (Tashlhiyt)	Afro-Asiatic	Africa	2,8898	1,8438	9,14
Burmese	Sino-Tibetan	East Asia	2,2901	3,1190	6,00
Chickasaw	Muskogean	America	2,5761	3,2281	5,70
Dinka	East Sudanic	Africa	1,9028	2,1022	13,70
Ende	Pahoturi (Papuan)	Australasia	2,1458	2,0140	10,21
Estonian	Uralic	Europe	2,6057	2,0349	9,56
Fataluku	Timor (Papuan)	Australasia	2,1053	2,1269	9,57
Georgian	Kartvelian	West Asia	2,3616	2,5286	7,78
Greek	Indo-European	Europe	2,1577	2,1346	11,56
Hausa	Afro-Asiatic	Africa	2,2979	1,6988	13,83

Hmong (Hmu)	Hmong-Mien	East Asia	2,1617	1,0637	15,70
Huehuetla (Tepehua)	Totonac	America	2,4413	2,8871	8,86
Hungarian	Uralic	Europe	2,2448	1,9200	10,00
Igbo	Niger-Congo	Africa	1,7115	1,9439	13,38
Ijo (Kalabari)	Niger-Congo	Africa	1,8882	1,8914	14,58
Irish	Indo-European	Europe	2,2809	1,3798	18,43
Italian	Indo-European	Europe	2,3085	1,7478	12,78
Japanese	Japonic	East Asia	1,9559	2,5506	9,89
Kabiye	Niger-Congo	Africa	1,9593	2,4286	10,11
Kazakh	Turkic	West Asia	2,5785	2,4778	11,25
Kera	Afro-Asiatic	Africa	2,3935	1,5650	9,32
Khongso	Sino-Tibetan	East Asia	2,5170	1,3968	21,00
Korean	Koreanic	East Asia	2,2952	2,7667	8,57
Kumiai	Yuman	America	2,9024	1,3443	7,63
Kunama	Kunaman	Africa	2,1337	3,0656	12,20
Lizu	Sino-Tibetan	East Asia	2,0930	1,9545	13,75
Mah Meri	Austro-Asiatic	East Asia	2,5385	1,6364	10,21
Makasar	Austronesian	Australasia	2,2873	2,4123	14,25
Malagasy	Austronesian	Africa	2,0435	2,1905	12,60
Malay	Austronesian	Australasia	2,3014	2,6795	9,75
Mambay	Niger-Congo	Africa	2,4023	1,3034	10,68
Mandarin	Sino-Tibetan	East Asia	2,6815	1,6020	9,80
Mapudungun	Araucanian	America	2,3841	2,0133	8,33
Mojeño	Arawakan	America	2,2065	2,3084	13,38
Mono	Niger-Congo	Africa	1,8674	1,5739	11,50
Nara	East Sudanic	Africa	2,3417	1,8426	9,82
Nen	Yam (Papuan)	Australasia	2,3021	2,3267	12,63
Paiute	Uto-Aztecan	America	2,1604	2,9444	4,50
Panjabi	Indo-European	West Asia	2,2644	1,5963	13,63
Persian	Indo-European	West Asia	2,4897	2,1319	10,11
Pitjantjatjara	Pama-Nyungan	Australasia	2,1792	2,9444	7,20
Qanjobal	Mayan	America	2,3750	2,4615	13,00
Qaqet	Baining (Papuan)	Australasia	2,2967	2,1329	9,29
Quichua (Ecuador)	Quechuan	America	2,1882	2,8830	8,55
Sahaptin	Penutian	America	2,4351	2,7018	7,13

Sandawe	Khoisan	Africa	2,1044	2,3038	8,78
Seenku	Niger-Congo	Africa	2,1078	1,3360	13,89
Seri	Hokan	America	2,4504	1,5414	14,27
Setswana	Niger-Congo	Africa	1,9188	1,6281	15,13
Shawi	Kawapanan	America	2,1312	3,4000	9,29
Shipibo	Panoan	America	1,7905	2,6079	11,95
Sumi	Sino-Tibetan	West Asia	1,8448	2,6667	12,43
Swedish	Indo-European	Europe	2,5917	1,5794	11,89
Tamambo	Austronesian	Australasia	2,1200	1,8072	11,86
Tamil	Dravidian	West Asia	2,1468	3,1899	8,78
Tausug	Austronesian	Australasia	2,4034	2,0877	9,50
Telugu	Dravidian	West Asia	2,1154	3,6000	8,13
Temne	Niger-Congo	Africa	2,1546	1,6560	11,36
Tepehuan	Uto-Aztecan	America	2,1326	2,2625	8,89
Tera	Afro-Asiatic	Africa	2,2390	1,6016	14,22
Thai	Tai-Kadai	East Asia	2,7746	1,3206	11,91
Tongan	Austronesian	Australasia	1,8566	1,9308	21,67
Trique	Oto-Manguean	America	2,2814	1,5182	18,33
Turkish	Turkic	West Asia	2,3552	2,7727	7,33
Ukrainian	Indo-European	Europe	2,5000	2,1667	12,00
Urarina	Urarinic	America	1,9349	2,9912	6,65
Vietnamese	Austro-Asiatic	East Asia	2,8547	1,0000	16,71
Zapotec (Isthmus)	Oto-Manguean	America	2,1234	1,7701	9,67
Average			**2,2491**	**2,1541**	**11,15**

Appendix 2. Exogenous Typological Variables

Language	Consonants	Vowels	Tones	Stress	Comp Syll	Cases
Abkhaz	59	2	1	1	1	2
Aingae	27	10	1	1	0	6
Albanian	29	7	1	0	1	4
Ambel	14	5	2	0	0	1
Amharic	27	7	1	1	0	2
Apache	33	8	3	0	0	1
Arabic	29	6	1	0	1	1
Arrernte	27	4	1	1	0	8
Bai	21	15	5	0	0	1
Basaa	30	14	4	0	0	1
Basque	23	5	1	1	1	10
Bengali	29	7	1	0	1	6
Berber	34	3	1	0	1	2
Burmese	34	9	4	0	0	8
Chickasaw	16	9	1	1	0	2
Dinka	20	7	4	0	0	1
Ende	19	7	1	0	0	6
Estonian	17	18	1	0	1	10
Fataluku	15	5	1	0	0	1
Georgian	28	5	1	1	1	6
Greek	18	5	1	1	1	3
Hausa	28	10	2	0	0	1
Hmong	32	8	8	0	0	1
Huehuetla	21	6	1	0	1	1
Hungarian	25	14	1	0	1	10
Igbo	26	8	3	0	0	1
Ijo	20	18	2	0	0	1
Irish	35	11	1	0	1	2
Italian	21	7	1	1	1	1
Japanese	16	5	2	1	0	8
Kabiye	21	9	2	0	0	1
Kazakh	20	11	1	0	0	6
Kera	24	6	3	0	1	1

Khongso	26	10	5	0	0	1
Korean	19	18	1	0	0	6
Kumiai	17	10	1	0	1	6
Kunama	22	10	3	0	0	6
Lizu	39	8	2	0	0	1
Mah Meri	30	19	2	0	0	3
Makasar	19	5	1	0	0	1
Malagasy	29	4	1	0	0	1
Malay	18	6	1	0	1	1
Mambay	25	10	2	0	0	1
Mandarin	19	6	4	0	0	1
Mapudungun	22	6	1	0	0	2
Mojeño	29	12	1	0	1	1
Mono	32	8	3	0	0	1
Nara	25	10	2	0	0	5
Nen	18	8	1	1	1	3
Paiute	17	11	1	0	0	5
Panjabi	27	17	3	0	1	2
Persian	23	6	1	1	1	2
Pitjantjatjara	17	6	1	0	0	10
Qanjobal	25	5	1	0	0	1
Qaqet	16	4	1	0	0	1
Quichua	23	3	1	0	0	8
Sahaptin	32	7	1	1	1	4
Sandawe	44	15	2	0	0	1
Seenku	20	12	4	0	0	1
Seri	18	8	1	0	1	1
Setswana	28	7	2	0	0	1
Shawi	12	4	1	1	0	6
Shipibo	15	8	1	1	0	6
Sumi	29	6	3	0	0	6
Swedish	16	17	1	1	1	2
Tamambo	16	5	1	0	0	1
Tamil	15	10	1	0	0	6
Tausug	17	3	1	0	0	1

Telugu	35	12	1	0	0	8
Temne	19	9	2	0	1	1
Tepehuan	12	5	2	0	0	1
Tera	35	11	3	0	0	1
Thai	21	9	5	0	0	1
Tongan	12	5	1	1	0	1
Trique	22	8	5	0	0	1
Turkish	22	8	1	0	0	6
Ukrainian	32	6	1	1	1	7
Urarina	13	13	2	0	0	1
Vietnamese	22	11	8	0	0	1
Zapotec	20	5	3	0	0	1
Average	**23,89**	**8,50**	**1,96**	**23%**	**31%**	**3,14**

Appendix 3. Exogenous Typological Variables (cont.)

Language	Genders	Inflections	Obj-Verb	Adj-Noun	Accusat	Isolating
Abkhaz	3	10	1	0	0	0
Aingae	1	6	1	1	1	0
Albanian	3	7	0	0	1	0
Ambel	1	6	0	0	1	1
Amharic	2	6	1	1	1	0
Apache	1	5	1	0	1	0
Arabic	2	6	0	0	1	0
Arrernte	1	4	1	0	0	0
Bai	1	2	0	1	0	1
Basaa	5	6	0	0	0	0
Basque	1	4	1	0	0	0
Bengali	2	2	1	1	1	0
Berber	2	6	0	0	1	0
Burmese	1	2	1	0	0	1
Chickasaw	1	6	1	0	0	0
Dinka	1	6	1	0	1	1
Ende	1	6	1	1	0	0
Estonian	1	2	0	1	1	0
Fataluku	1	4	1	0	0	1
Georgian	1	8	1	1	1	0
Greek	3	4	0	1	1	0
Hausa	2	6	0	1	0	1
Hmong	1	2	0	0	0	1
Huehuetla	1	4	0	1	1	0
Hungarian	1	4	0	1	1	0
Igbo	1	6	0	0	0	1
Ijo	1	6	1	1	1	1
Irish	2	2	0	0	1	0
Italian	2	4	0	0	1	0
Japanese	1	4	1	1	0	0
Kabiye	1	2	0	0	0	1
Kazakh	1	6	1	1	1	0
Kera	2	6	0	0	1	1

Khongso	1	1	1	1	0	1
Korean	1	6	1	1	0	0
Kumiai	1	6	1	0	1	0
Kunama	2	4	1	0	1	0
Lizu	1	3	1	0	0	1
Mah Meri	1	1	0	0	0	1
Makasar	1	5	0	0	0	1
Malagasy	1	4	0	0	1	0
Malay	1	4	0	0	1	1
Mambay	1	1	0	0	0	0
Mandarin	1	1	0	1	0	1
Mapudungun	1	8	0	1	0	0
Mojeño	1	6	0	0	0	0
Mono	5	6	1	0	0	0
Nara	2	4	1	0	1	1
Nen	1	10	1	0	0	0
Paiute	1	4	1	1	1	0
Panjabi	2	3	1	1	1	0
Persian	1	4	1	0	1	0
Pitjantjatjara	1	4	1	0	1	0
Qanjobal	1	4	0	1	0	0
Qaqet	8	3	0	0	0	0
Quichua	1	8	1	1	1	0
Sahaptin	1	10	0	1	0	0
Sandawe	5	8	1	0	1	1
Seenku	1	2	1	0	0	1
Seri	1	5	1	1	0	0
Setswana	5	4	0	0	1	0
Shawi	1	6	1	1	0	0
Shipibo	1	6	1	1	0	0
Sumi	1	4	1	0	1	0
Swedish	3	2	0	1	1	0
Tamambo	1	6	0	0	1	1
Tamil	3	2	1	1	1	0
Tausug	1	4	0	0	0	1

Telugu	3	2	1	1	1	0
Temne	5	2	0	0	1	0
Tepehuan	1	4	0	1	1	0
Tera	1	2	0	0	0	1
Thai	1	2	0	0	0	1
Tongan	1	6	0	0	0	1
Trique	1	6	0	0	1	0
Turkish	1	6	1	1	1	0
Ukrainian	3	4	0	1	1	0
Urarina	1	8	1	0	0	0
Vietnamese	1	1	0	0	0	1
Zapotec	1	8	0	0	1	0
Average	**1,65**	**4,63**	**49%**	**40%**	**53%**	**33%**