UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

A STRUCTURE-CONDUCT-PERFORMANCE APPROACH TO LANGUAGE COMPLEXITY TRADE-OFFS

Germán Coloma

Febrero 2026
Nro. 918

https://ucema.edu.ar/publicaciones/doc_trabajo.php
UCEMA: Av. Córdoba 374, C1054AAP Buenos Aires, Argentina
ISSN 1668-4575 (impreso), ISSN 1668-4583 (en línea)
Editor: Jorge M. Streb; Coordinador del Departamento de Investigaciones: Maximiliano Ivickas

A Structure-Conduct-Performance Approach to Language Complexity Trade-Offs

Germán Coloma [*]

Abstract

In this paper, we present an approach to relate typological measures of language complexity (based on the grammars of different languages) with empirical measures of that complexity (based on actual texts). It is known as the "structure-conduct-performance paradigm", and we have taken it from the field of industrial economics. Using a sample of 45 languages for which we have the same text, we apply that approach to capture some relationships that go from phylogenetic, geographic, demographic and sociological characteristics of languages (structural variables) towards some typological variables that determine measures of phonological and morphological complexity, and then have an impact on two corpus-based language ratios (phonemes per word and words per clause). Our results, based on correlation and regression analyses, show some important trade-offs between the typological measures, between the language ratios, and between both sets of variables. Those results are also robust to extending the number of languages in the sample to 81 observations, and to dividing that extended sample into sub-samples.

Keywords: structure-conduct-performance, complexity trade-offs, correlation, language ratios.

1. Introduction

The existence of complexity trade-offs across languages has to do with the idea that a language which is more complex in a certain dimension (e.g., in its morphology) must be simpler in another dimension (e.g., in its phonology or syntax). In general, the literature about language complexity trade-offs has positive results (i.e., it finds evidence in favor of the existence of those trade-offs) when it deals with empirical measures of complexity (i.e., measures derived from actual texts), but it has dubious or sometimes negative results when it uses theoretical or typological measures (i.e., measures derived from grammatical descriptions of the languages). This can be due to different causes, which may be related to the way in which complexity measures are computed.

In this paper, we will try to reconcile those conclusions, using an approach that was originally developed in the field of empirical industrial economics and is known as

[*] CEMA University; Av. Cordoba 374, Buenos Aires, C1054AAP, Argentina. Telephone: 54-11-6314-3000. E-mail: gcoloma@cema.edu.ar. The author's viewpoints do not necessarily represent the position of the University.

the "structure-conduct-performance paradigm". This basically implies that there are some conditions that have to do with the structure of a problem, which influence the behavior (conduct) of the agents that make decisions in that problem. Those decisions, in turn, determine the outcomes (or the "performance") that can be observed as a consequence of the problem's solution.

In industrial economics, for example, the typical structural conditions have to do with the degree of market concentration (i.e., if a market has a few large firms or many small firms), the existence of product differentiation, the presence of entry barriers, etc. Conduct, conversely, has to do with decisions that firms make about variables such as price, quantity, product quality and advertising. Finally, performance is typically measured by the profit rates or the profit margins of the participating firms.[1]

The structure-conduct-performance approach, however, can also be applied to other environments. In the context of languages, for example, we could consider that the structural conditions in which a language develops have to do with some geographic characteristics (e.g., the region in which each language originated or is spoken), some phylogenetic characteristics (e.g., the language family to which it belongs) some demographic characteristics (e.g., the number of speakers) and some sociological characteristics (e.g., if a language is spoken in different countries or is widely used as a second language).

The variables related to the grammar of a language, conversely, can be seen as a manifestation of conduct. For example, a language may have more or fewer phonemes, which can be consonants or vowels. It may also have only one or several contrastive tones, and its verbs can have a single form or multiple inflections based on tense, aspect, person, etc. Moreover, its nouns can be invariant, or else have many variations based on gender, number or case declensions. All those variables can be used to build some complexity measures, both at the phonological and at the morphological level.

Finally, the performance of a language can be assessed using empirical (corpus-based) measures derived from actual texts. Those measures may be ratios that are calculated using those texts, such as phonemes per word or words per clause.

In previous work (Coloma, 2016, 2017a, 2022), we analyzed the possible existence of complexity trade-offs using empirical measures for samples of different languages for which we had the same text. Besides, we also tried to apply similar

[1] For a good explanation of the logic of the structure-conduct-performance paradigm in industrial economics, see Kumar & Choudhary (2024) or Perloff, Karp & Golan (2007), chapter 2.

procedures but focusing on "conduct measures" of language complexity, given by typological or theoretical variables (see, for example, Coloma, 2017b, 2024). In this paper, we will try to combine both strategies using an empirical approach which assumes that there are several structural variables (phylogenetic, geographic, demographic and sociological) that influence some typological variables (phonological and morphological), which in turn determine performance (i.e., the values of the different language ratios). The trade-offs would occur both at the level of the typological (conduct) measures and at the level of performance, and we will see that the inclusion of structural variables, which have an influence on them, helps to increase the significance of some trade-offs.

The sample on which we will conduct our experiment has 45 languages from different families and regions, and it also exhibits considerable typological variation in 6 measurable grammatical characteristics (size of the consonant phoneme inventory, size of the vowel phoneme inventory, number of contrastive tones, number of cases, number of genders, and number of inflectional categories of the verbs). For those languages, we have the same text (the fable "The North Wind and the Sun"), whose English version has 113 words.[2]

The way in which we will try to discover the possible existence of language complexity trade-offs is essentially based on the computation of correlation coefficients between the different variables. At the level of the conduct variables, for example, we will see if there is a negative correlation between measures of phonological and morphological complexity. Correspondingly, at the level of the performance variables, we will detect the possible correlation between phonemes per word and words per clause, and we will also try to find some correlations between the typological measures and the empirical language ratios (e.g., phonological complexity vs. phonemes per word, or morphological complexity vs. words per clause).

In order to implement the structure-conduct-performance approach, we will see if the structural variables (which in our exercise will be categorical variables related to world regions, language families, language size, and language use) have some influence on conduct variables, and if those variables have an influence on performance variables. This will imply recalculations of the correlation coefficients, using previous regression

[2] Most versions of "The North Wind and the Sun" that we use here are taken from publications of the International Phonetic Association (IPA), such as IPA (1949), IPA (1999), or the "Illustrations of the IPA" published in the *Journal of the International Phonetic Association*. For some examples, see appendix 6.

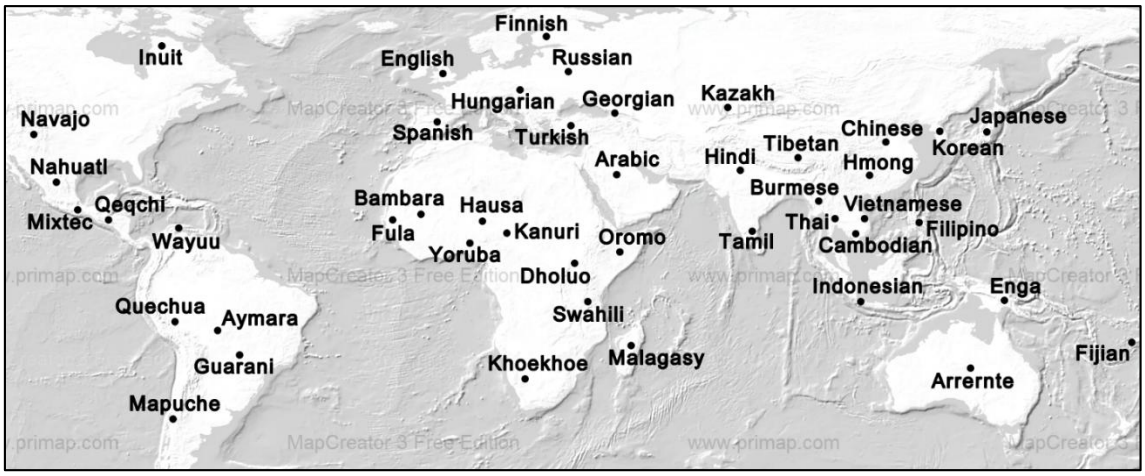
analyses in which conduct variables are related to structural variables, and performance variables are related to conduct variables.

The organization of this paper is as follows. In section 2 we will describe the sample under analysis. In section 3 we will see some relationships between the variables of our sample, computing several standard correlation coefficients that might signal the possible existence of language complexity trade-offs. In section 4, in turn, we will describe the structure-conduct-performance model to be applied to our data. The results obtained will be shown in section 5, and in section 6 they will be compared to those that arise when we use an extended language sample, and four different sub-samples. Finally, in section 7, we will present the main conclusions of the whole paper.

2. Description of the data

As we mentioned in the introduction, the sample that we will use consists of 45 languages (see appendix 1), which belong to 9 different world regions. From each region, we have chosen 5 languages, which are Inuit, Navajo, Nahuatl, Mixtec and Qeqchi (North America); Wayuu, Quechua, Aymara, Guarani and Mapuche (South America); Fula, Bambara, Yoruba, Hausa and Kanuri (West Africa); Oromo, Dholuo, Swahili, Malagasy and Khoekhoe (East Africa); Spanish, English, Hungarian, Finnish and Russian (Europe); Georgian, Turkish, Arabic, Hindi and Tamil (West Asia); Kazakh, Tibetan, Chinese, Korean and Japanese (North Asia); Hmong, Burmese, Thai, Cambodian and Vietnamese (South Asia); and Filipino, Indonesian, Arrernte, Enga and Fijian (Australasia). The approximate location of those languages can be seen on the map of figure 1.

Figure 1: Location of the languages included in the sample



Our languages belong to 28 different families but, in the case of the most important ones, we have included more than one language. In fact, our sample has 4 Indo-European languages (English, Spanish, Russian and Hindi), 4 Niger-Congo languages (Fula, Bambara, Yoruba and Swahili), 4 Austronesian languages (Malagasy, Indonesian, Filipino and Fijian), 3 Afro-Asiatic languages (Hausa, Oromo and Arabic), 3 Sino-Tibetan languages (Chinese, Tibetan and Burmese), 2 Turkic languages (Turkish and Kazakh), 2 Uralic languages (Hungarian and Finnish), 2 Austro-Asiatic languages (Cambodian and Vietnamese) and 2 Nilo-Saharan languages (Kanuri and Dholuo), but each one belongs to a different "language genus".[3]

To choose the languages that make up the sample, we tried to include the most important genera within each family, and the most important language (in terms of the number of speakers) within each genus. Due to that, we have 8 languages with more than 100 million native speakers (Chinese, English, Spanish, Arabic, Hindi, Russian, Japanese and Indonesian), that we will call "major languages". We also have 17 languages with more than 10 million native speakers but less than 100 million (Burmese, Cambodian, Filipino, Fula, Hausa, Hungarian, Kanuri, Kazakh, Korean, Malagasy, Oromo, Swahili, Tamil, Thai, Turkish, Vietnamese and Yoruba), which are considered to be "large languages".

Eleven other languages in our sample have between one million and 10 million native speakers (Aymara, Bambara, Dholuo, Finnish, Georgian, Guarani, Hmong, Nahuatl, Qeqchi, Quechua and Tibetan), and they are classified as "medium-size languages". Finally, the remaining 9 languages (Arrernte, Enga, Fijian, Inuit, Khoekhoe, Mapuche, Mixtec, Navajo and Wayuu) have less than one million native speakers, and they are therefore considered to be "small languages".

Following an idea that appears in Chen et al. (2024), we have also classified languages according to a sociological criterion related to their use. According to it, 15 languages in our sample (Arabic, Bambara, Chinese, English, Filipino, Fula, Hausa, Hindi, Indonesian, Kanuri, Quechua, Russian, Spanish, Swahili and Thai) are "exoteric", while the remaining 30 languages are considered to be basically local. The exoteric languages are the ones spoken as first languages in several countries, or the ones that have a substantial proportion of second-language speakers.

The typological information that we have collected for each language (see

[3] For example, the Indo-European languages included in the database belong to the following genera: Germanic (English), Indic (Hindi), Romance (Spanish) and Slavic (Russian).

appendix 2) corresponds to the six grammatical variables mentioned in the introduction (number of consonants, number of vowels, number of tones, number of cases, number of genders, and number of inflectional categories of the verbs).[4] Three of them (consonants, vowels and tones) are phonological variables, and the other three (cases, genders and inflections) are morphological variables.

Looking at the whole database, we find that the size of the consonant inventories of the included languages ranges from a minimum of 13 (Finnish) to a maximum of 58 (Hmong), while the number of vowel phonemes varies between 3 (Quechua) and 20 (Cambodian). Twenty-seven languages have only one distinctive tone, but others have many more, and the language with the largest number of tone distinctions in our sample (8) is Vietnamese. An analogous situation occurs with the number of cases. While 25 languages do not distinguish between cases (and therefore have only one case declension), others have several distinctions, and the maximum number of different cases (10) corresponds to two languages: Finnish and Hungarian.

Thirty-one languages in the database do not distinguish between genders or classes of nouns, but the remaining 14 languages do. Some of them have a relatively simple two-way gender distinction (e.g., Arabic, Filipino, Hausa, Hindi, Oromo, Spanish, Wayuu), but the largest number of grammatical noun classes in our sample is 8 and corresponds to the Fula language (which classifies nouns in categories such as humans, animals, instruments, natural forces, etc.).

Concerning verbs, we see that three languages (Cambodian, Chinese and Vietnamese) have a single form for each verb, while the maximum number of inflectional categories in our sample is equal to 8, and corresponds to Aymara, Georgian, Kanuri, Mapuche, Nahuatl and Quechua.

Using the values of our typological variables, we have built a measure of phonological complexity (*Phoncomp*) and a measure of morphological complexity (*Morphcomp*). The formulae that define those measures are the following:

$$Phoncomp = Consonants + Vowels*Tones \quad (1)\ ;$$

$$Morphcomp = Cases + Genders + Inflections \quad (2)\ ;$$

and these variables also display considerable variation across languages (see appendix 3). The phonological complexity measure, for example, ranges from a minimum of 20 (Inuit)

[4] In the majority of the cases, this information has been taken from Dryer & Haspelmath (2013).

to a maximum of 121 (Hmong). Similarly, the languages with a lowest morphological complexity value (3) are Vietnamese, Cambodian and Chinese, while the language with the highest morphological complexity in our sample (17) is Quechua.

Additionally, the data that we have from our text ("The North Wind and the Sun") consists of the number of clauses, words and phonemes of that text in each language. With those numbers, we calculated two ratios (phonemes/words and words/clauses), which can also be seen as corpus-based measures of language complexity. The numbers from these measures vary considerably when we compare the different languages, as can be seen in appendix 3. The phoneme/word ratio goes from a minimum of 2.85 (Vietnamese) to a maximum of 9.95 (Inuit), while the word/clause ratio goes from a minimum of 6.08 (Arrernte) to a maximum of 16.86 (Mixtec).

3. Correlation analysis

The data series described in the previous section show certain correlations between them. In table 1, for example, we see the standard (Pearson) correlation coefficients for our two language ratios (*Phonword* and *Wordclaus*) and our two typological complexity measures (*Phoncomp* and *Morphcomp*).

Table 1: Standard correlation coefficients between complexity measures

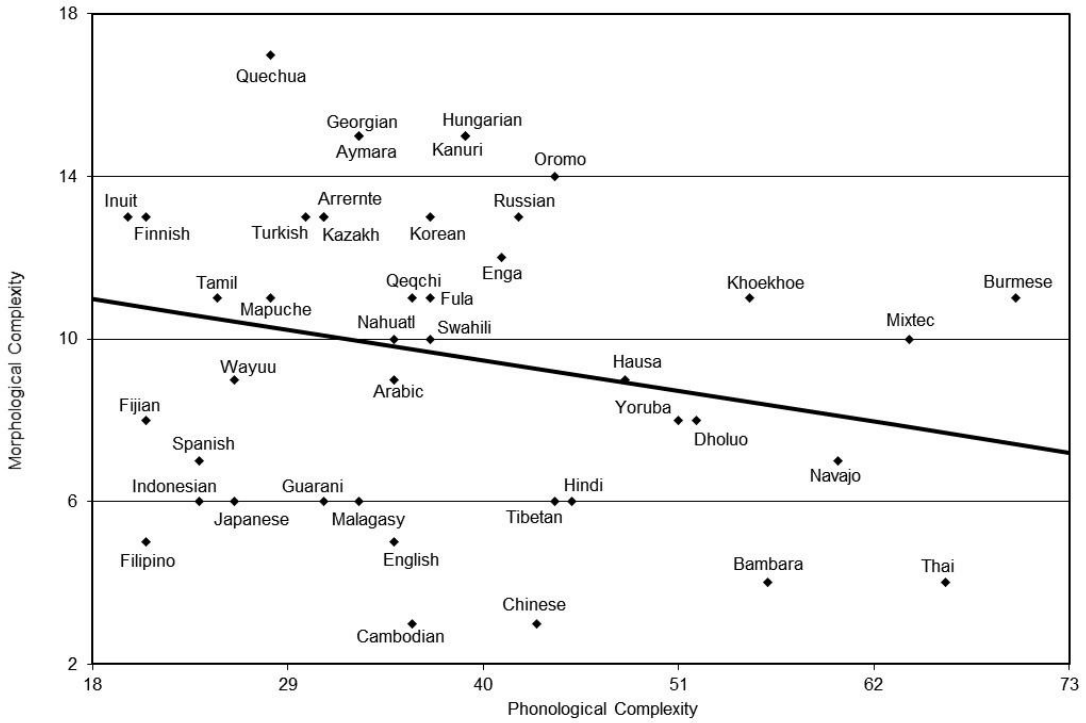
Variable	Phonword	Wordclaus	Phoncomp	Morphcomp
Phonemes per word	1.0000			
Words per clause	-0.7451	1.0000		
Phonological complexity	-0.4851	0.5006	1.0000	
Morphological complexity	0.5815	-0.5481	-0.3641	1.0000

As can be observed, there are several negative correlation coefficients that display considerably large values, and these can be seen as signs of language complexity trade-offs. One possible way to analyze if the absolute value of a negative correlation coefficient is high enough to signal a meaningful trade-off is to calculate its statistical significance. As we are working with a sample of 45 observations, we can consider that a statistically significant correlation occurs if the corresponding coefficient is above 0.294 in absolute value. This is because the probability for that correlation to be null is below 5%. Moreover, if the coefficient is above 0.38, then that probability will be below 1% (and we can say that there is a "highly significant correlation").[5]

[5] These numbers have been calculated using a "t-statistic" formula, that relates the value of the correlation

For example, table 1 shows that the correlation between phonemes per word and words per clause (-0.7451) is highly significant, as is the correlation between morphological complexity and words per clause (-0.5481). There is also one very high positive correlation coefficient, which is the one between phonemes per word and morphological complexity (0.5815).

Figure 2: Correlation between phonological and morphological complexity



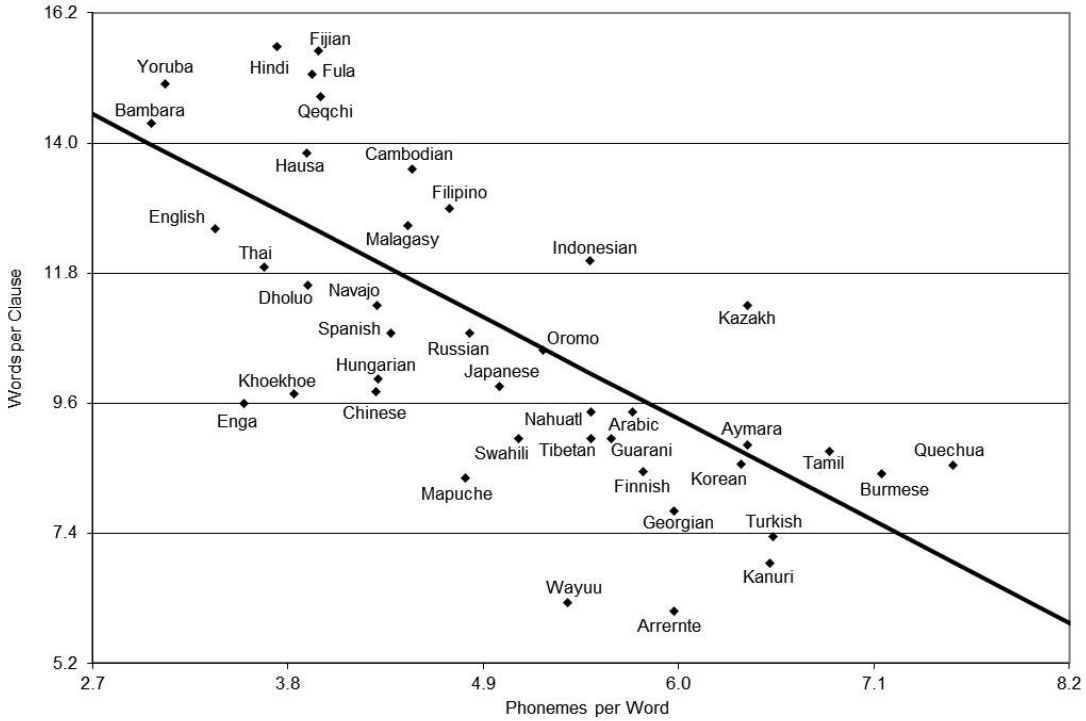
Phonological complexity and morphological complexity are negatively correlated, but the corresponding correlation coefficient (-0.3641) is not as large as those mentioned in the previous paragraph. Moreover, that coefficient is statistically significant at a 5% probability level (because it is greater than 0.294 in absolute value), but not at a 1% probability level (since its absolute value is less than 0.38). This can be seen in figure 2, in which we show the relationship between these two variables in a diagram where each language is depicted as a black rhomb, and correlation is represented by a thick straight trend line with a relatively small negative slope.

In figure 3, conversely, we see that the correlation between phonemes per word

coefficient and the number of observations in the sample (which determines the "degrees of freedom" of the estimation). This statistic has a certain distribution, which indicates the probability that the corresponding correlation coefficient is actually equal to zero. For a complete explanation of this, see Bonamente (2022), chapter 14.

and words per clause (i.e., between our two empirical language ratios) is much stronger, since the corresponding trend line is steeper and, on average, languages are closer to that line.

Figure 3: Correlation between phonemes per word and words per clause



This difference between the two represented correlations can be related to other results that appear in the literature. As we mentioned in the introduction, it is relatively common to find high negative correlations between corpus-based measures of complexity (as can be the language ratios that we have calculated),[6] while the correlations between typological measures of complexity tend to be much smaller (and, sometimes, insignificant).[7]

It is, of course, possible to measure some complexity trade-offs using combinations of typological and corpus-based measures, whose correlations can also be interpreted as signs of those trade-offs. Indeed, the correlation between *Phoncomp* and *Phonword*, which is reported on table 1 (-0.4851), can be seen as an indirect measure of the relationship between phonological and morphological complexity, while the correlation between *Morphcomp* and *Wordclaus* (-0.5481) can be seen as indicative of a

[6] See, for example, Fenk-Oczlon & Fenk (2008), Oh & Pellegrino (2023) or Bentz et al. (2023).
[7] See, for example, Shosted (2006), Nichols (2009), Shcherbakova et al. (2023) or Benítez, Chen & Gil (2024).

trade-off between morphological and syntactic complexity.

Some of those trade-offs, however, could be hidden behind other factors that occur simultaneously, which implies that standard correlation coefficients (like those shown in table 1) might not be suitable to detect them. It is also possible that some coefficients are relatively high because of the existence of "spurious correlations" (i.e., correlations that have to do with other unseen factors, which are the ones that are truly correlated with the variables under analysis). In that case, it may occur that correlation analysis points out some trade-offs that are not real.[8]

In order to solve the problems of correlation analysis, it is possible to use different strategies. In the next section we will present a model that assumes certain interactions between groups of variables, and those interactions will have specific representations in terms of regression equations. Using those representations, we will calculate new correlation coefficients, and those coefficients will be useful for detecting, confirming or ruling out the existence of complexity trade-offs between the different language variables.

4. A structure-conduct-performance approach

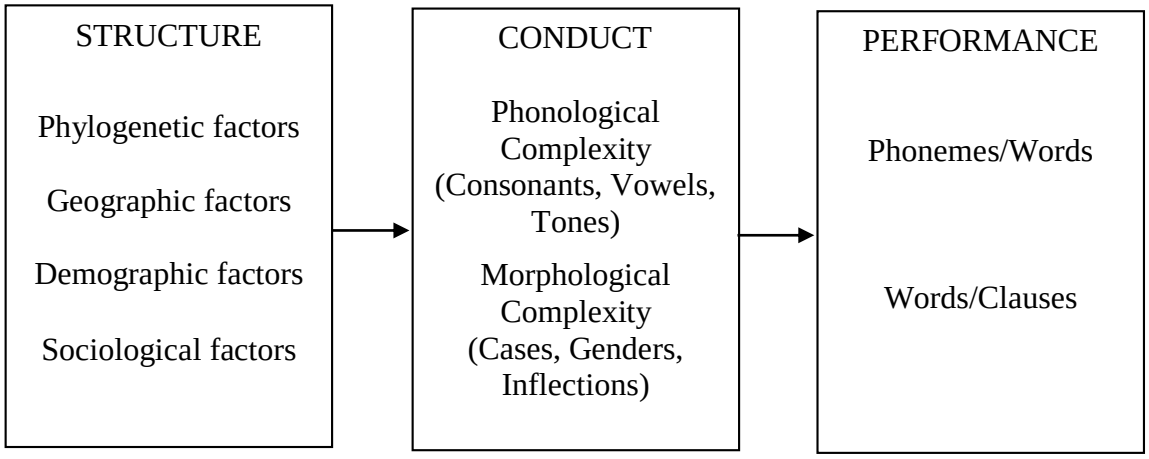
The structure-conduct-performance (SCP) paradigm, outlined in the introduction of this paper, assumes a direction of causality from structural variables (in our case, phylogenetic, geographic, demographic and sociological factors) towards conduct variables (i.e., typological complexity measures, based on the number of consonants, vowels, tones, cases, genders and inflections in each language) to performance variables (i.e., empirical or corpus-based language ratios, such as phonemes/words and words/clauses).

All this is depicted in figure 4, in which we show the idea that the structural variables are determined outside the system, but exert an influence on the conduct variables. That influence can be measured through the difference between the typological characteristics of languages associated with various regions, families or numbers of speakers. The last element of the model is the relationship between typological and empirical complexity, for which we will assume that it is the first of them the one that determines the latter. This is because we can presume that the people who create oral or written texts in a language take the grammar of that language as given, and therefore it is

[8] For an interesting analysis of the relationship between language trade-offs and negative correlation coefficients, see Levshina (2020).

the features of that grammar that influence the characteristics of the texts (and not the other way round).

Figure 4: Structure-conduct-performance diagram



This SCP model can also be represented through equations, which will be used to run a series of regressions for each module. The relationship between structure and conduct, for example, can be written like this:

*Phoncomp = c(1)*Europe +c(2)*Westafrica +c(3)*Eastafrica +c(4)*Northamerica +c(5)*Southamerica +c(6)*Westasia +c(7)*Southasia +c(8)*Northasia +c(9)*Australasia +c(10)*Indoeuro +c(11)*Afroasiatic +c(12)*Nigercongo +c(13)*Austronesian +c(14)*Turkic +c(15)*Sinotibetan +c(16)*Austroasiatic +c(17)*Nilosaharan +c(18)*Uralic +c(19)*Major +c(20)*Large +c(21)*Medium +c(22)*Exoteric* (3) ;

*Morphcomp = c(23)*Europe +c(24)*Westafrica +c(25)*Eastafrica +c(26)*Northamerica +c(27)*Southamerica +c(28)*Westasia +c(29)*Southasia +c(30)*Northasia +c(31)*Australasia +c(32)*Indoeuro +c(33)*Afroasiatic +c(34)*Nigercongo +c(35)*Austronesian +c(36)*Turkic +c(37)*Sinotibetan +c(38)*Austroasiatic +c(39)*Nilosaharan +c(40)*Uralic +c(41)*Major +c(42)*Large +c(43)*Medium +c(44)*Exoteric* (4) ;

where *Europe, Westafrica, Eastafrica, Northamerica, Southamerica, Westasia, Southasia, Northasia* and *Australasia* are binary variables (whose values can either be zero or one) that represent the different world regions; *Indoeuro, Afroasiatic, Nigercongo, Austronesian, Turkic, Sinotibetan, Austroasiatic, Nilosaharan* and *Uralic* are binary variables that represent the families with two or more languages in the sample; *Major, Large* and *Medium* are binary variables that divide the sample in demographic

categories;[9] and *Exoteric* is a sociological binary variable related to language use.

As a result of the regressions performed using this system, we can obtain fitted values for our typological complexity measures. These fitted values generate two variables that we will label as *Phonĉomp* and *Morphĉomp*. Here the symbol "^" represents the idea that these are "instrumental variables", built as the outcome of a regression against other exogenous variables.[10]

If we now model the relationship between conduct and performance, we can use equations like these ones:

$$Phonword = c(1) + c(2)*Consonants + c(3)*Vowels + c(4)*Tones + c(5)*Cases + c(6)*Genders + c(7)*Inflections \quad (5)\ ;$$

$$Wordclaus = c(8) + c(9)*Consonants + c(10)*Vowels + c(11)*Tones + c(12)*Cases + c(13)*Genders + c(14)*Inflections \quad (6)\ ;$$

and the regressions performed using this system lead to the creation of instrumental variables for the corpus-based measures (*Phonŵord* and *Wordĉlaus*). Note that in this case we use the original typological variables (*Consonants, Vowels, Tones, Cases, Genders, Inflections*) and not the derived complexity measures (*Phoncomp, Morphcomp*). This is because using a larger set of variables produces a better fit for the estimated instrumental variables, and it therefore increases the precision of the corresponding estimation.

The next step of the procedure is to relate the instrumental conduct variables (*Phonĉomp* and *Morphĉomp*) with the instrumental performance variables (*Phonŵord* and *Wordĉlaus*). This implies calculating new correlation coefficients, that we will call "SCP coefficients". With those coefficients, we will check if there are significant trade-offs between the typological variables, between the empirical variables, and between typological and empirical variables.

5. Estimation results

If we apply the model introduced in section 4 to the database described in section 2, we obtain different results that can be expressed as regression coefficients. For

[9] In our sample, those demographic categories are four: major languages (for which *Major* = 1, *Large* = 0, and *Medium* = 0), large languages (for which *Major* = 0, *Large* = 1, and *Medium* = 0), medium languages (for which *Major* = 0, *Large* = 0, and *Medium* = 1) and small languages (for which *Major* = 0, *Large* = 0, and *Medium* = 0).

[10] For a good explanation of the logic of instrumental variables in statistical analysis, see Angrist & Pischke (2009), chapter 4.

example, if we run the least-square regressions that appear in equations 3 and 4, we obtain the following output:

$$\hat{Phoncomp} = 26.6238*Europe + 58.3476*Westafrica + 53.4602*Eastafrica + 43.2044*Northamerica + 30.0296*Southamerica + 37.9571*Westasia + 96.4557*Southasia + 45.2219*Northasia + 36.3709*Australasia + 16.0734*Indoeuro + 5.2755*Afroasiatic + 0.31799*Nigercongo - 6.2454*Austronesian - 2.295*Sinotibetan + 2.3835*Turkic - 9.9827*Austroasiatic - 2.1044*Nilosaharan + 10.3682*Uralic - 6.4155*Major - 13.473*Large - 0.51096*Medium - 2.615*Exoteric \quad (7)\ ;$$

$$\hat{Morphcomp} = 13.0685*Europe + 9.1326*Westafrica + 10.0219*Eastafrica + 10.0219*Northamerica + 11.3846*Southamerica + 10.7352*Westasia + 4.3227*Southasia + 7.4797*Northasia + 12.6571*Australasia - 3.0723*Indoeuro - 0.73613*Afroasiatic - 3.3622*Nigercongo - 6.828*Austronesian - 0.39494* Sinotibetan + 0.77225*Turkic - 4.4429*Austroasiatic + 0.26941*Nilosaharan - 0.85137*Uralic - 1.4039*Major + 3.1203* Large + 0.44537*Medium - 0.25898*Exoteric \quad (8)\ .$$

Similarly, running least-square regressions for equations 5 and 6 produces these results:

$$\hat{Phonword} = 5.5296 - 0.01264*Consonants - 0.06939*Vowels - 0.2411*Tones + 0.23325*Cases - 0.10945*Genders + 0.05741*Inflections \quad (9)\ ;$$

$$\hat{Wordclaus} = 9.8134 + 0.03887*Consonants + 0.14265*Vowels + 0.412*Tones - 0.39043*Cases + 0.40605*Genders - 0.2919*Inflections \quad (10)\ .$$

With all these outcomes, it is possible to compute values for *Phonĉomp*, *Morphĉomp*, *Phonŵord* and *Wordĉlaus*, which correspond to each observation in the sample. In turn, those values can be used to calculate new correlation coefficients for the different pairs of variables, which are the ones that appear on table 2. These will be the SCP coefficients that we will interpret as possible signals of language complexity trade-offs.

Table 2: Correlation coefficients under the SCP approach

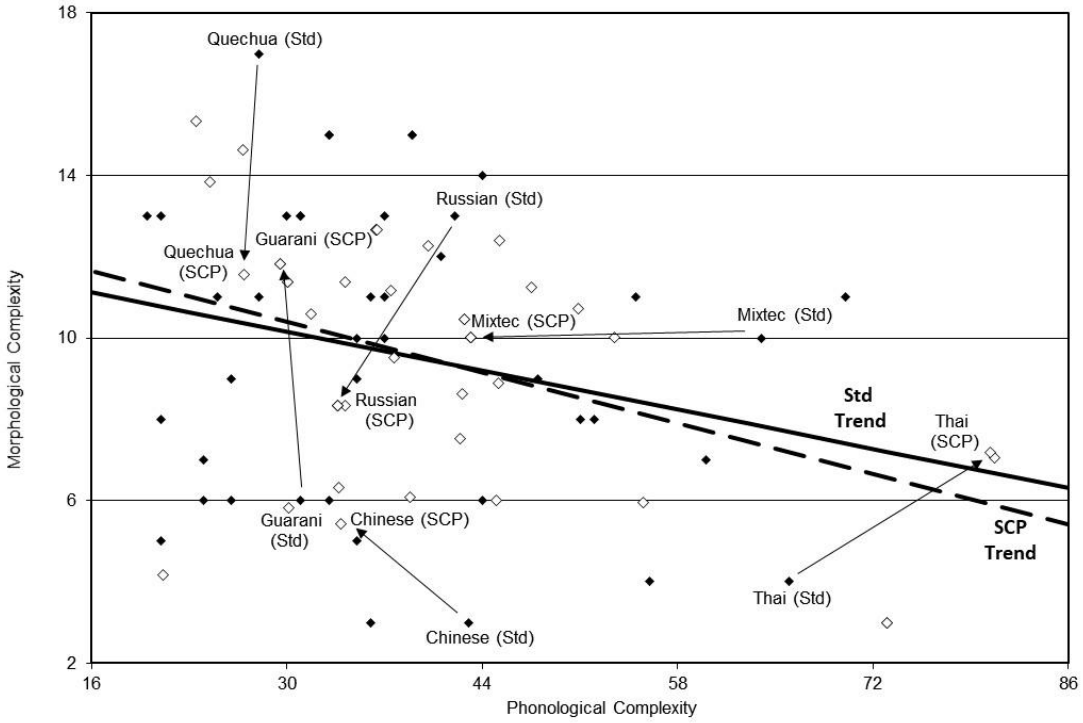
Variable	Phonword	Wordclaus	Phoncomp	Morphcomp
Phonemes per word	1.0000			
Words per clause	-0.9673	1.0000		
Phonological complexity	-0.5660	0.5694	1.0000	
Morphological complexity	0.6256	-0.6185	-0.4897	1.0000

As we see, the new correlation coefficients have the same signs (either positive or negative) as the corresponding standard correlation coefficients reported in table 1. Besides, the absolute values of these coefficients are greater than the corresponding

standard coefficients. Moreover, all of them, including the correlation coefficient that relates phonological and morphological complexity, are now statistically significant at a 1% probability level (since they are all greater than 0.38 in absolute value).

The main differences between standard correlation and correlation under the SCP approach can be illustrated by diagrams like the one shown on figure 5, in which we have depicted the relationship between phonological and morphological complexity. Here the black rhombs represent the original observations of our sample, and are identical to those that appear in figure 2. The white rhombs, conversely, come from the fitted values of equations 7 and 8, and show the values of *Phonĉomp* and *Morphĉomp* that correspond to each of the 45 languages of the sample.

Figure 5: Standard Correlation and SCP Correlation



As we see, on average, this last set of values is closer to the corresponding trend line ("SCP Trend", drawn as a dashed line) than what the original observations are in relationship with the original trend line ("Std Trend", drawn as a thick line). That is particularly clear in the case of some languages that are outliers in the original sample (e.g., Chinese, Guarani, Mixtec, Quechua, Russian, Thai), for which we have depicted arrow lines that connect the original observations with the corresponding SCP fitted values. Moreover, because of the adjustment made by the regression procedure, the correlation coefficient under the structure-conduct-performance approach (-0.4897)

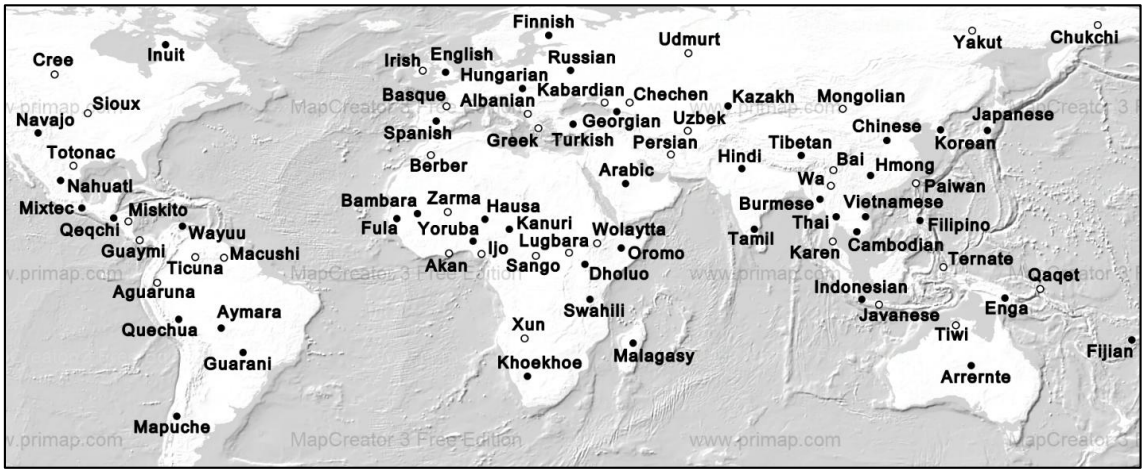
becomes greater than the standard coefficient (-0.3641), and that difference also explains why the SCP trend line in figure 5 is steeper than the original trend line.

6. Robustness checks with an extended sample

The outcomes of the previous section can be subjected to some robustness checks. In this section we will make a few of them, using an extended sample of languages. That sample consists of the original 45 observations, plus 36 additional ones. The additional observations correspond to four languages from each of the nine world regions defined in section 2.

The list of additional languages is the following: Cree, Sioux, Totonac and Miskito (North America); Guaymi, Aguaruna, Ticuna and Macushi (South America); Berber, Zarma, Akan and Ijo (West Africa); Wolaytta, Lugbara, Sango and Xun (East Africa); Irish, Basque, Albanian and Greek (Europe); Kabardian, Chechen, Uzbek and Persian (West Asia); Udmurt, Mongolian, Yakut and Chukchi (North Asia); Paiwan, Bai, Wa and Karen (South Asia); and Javanese, Tiwi, Ternate and Qaqet (Australasia). Although many of these languages belong to families that are already included in the original sample (Indo-European, Niger-Congo, Afro-Asiatic, Austronesian, etc.), all of them are from different genera. In the map of figure 6, the additional languages are depicted as white circles, while the languages from the original sample are depicted as black circles.[11]

Figure 6: Languages from the extended sample



[11] The characteristics associated to the additional languages are summarized in the tables of appendix 4. In that appendix we can see, for example, that in the extended sample there is one language with 94 consonants (Xun), another one with only 2 vowels (Kabardian), and another one with 10 tones (Ticuna).

In table 3, we can see the main results of applying the structure-conduct-performance approach to our extended sample of 81 languages. The first half of the table shows the standard correlation coefficients, while the second half shows the SCP correlation coefficients.[12]

Table 3: Correlation coefficients for the extended sample of 81 languages

Variable	Phonword	Wordclaus	Phoncomp	Morphcomp
Standard correlation				
Phonemes per word	1.0000			
Words per clause	-0.7243	1.0000		
Phonological complexity	-0.5205	0.4559	1.0000	
Morphological complexity	0.5397	-0.4150	-0.2986	1.0000
SCP correlation				
Phonemes per word	1.0000			
Words per clause	-0.9648	1.0000		
Phonological complexity	-0.5691	0.5416	1.0000	
Morphological complexity	0.6204	-0.5567	-0.4723	1.0000

The figures of table 3 show, once again, that all the estimated coefficients have the same signs as those found for the original sample, and this holds for both the standard coefficients and the SCP coefficients. We can also see that the SCP correlation coefficients display larger absolute values than the standard coefficients, just as it happens in the original sample. Moreover, when we apply the SCP approach, all correlation coefficients are significant at a 1% probability level.[13]

One problem in the comparison between the results obtained in the previous section and the ones that arise when we use an extended sample of languages is that the number of observations is not the same. This makes comparing correlation coefficients inappropriate, since the statistical significance of those coefficients is not equivalent in a sample of 45 observations as in a sample of 81 observations. Nevertheless, we can use the extended sample to make several sub-samples, and those sub-samples can have the same number of observations as the original sample. Moreover, we can build sub-samples using different criteria, to see if our results are robust to changes in the characteristics of the languages included in each sub-sample. That is what we will do in the remaining part of this section, in which we will work with four different sub-samples.

Our first sub-sample consists of the 36 additional observations that we have

[12] The regression coefficients that were used to compute the values for the instrumental variables are reported in the tables of appendix 5.

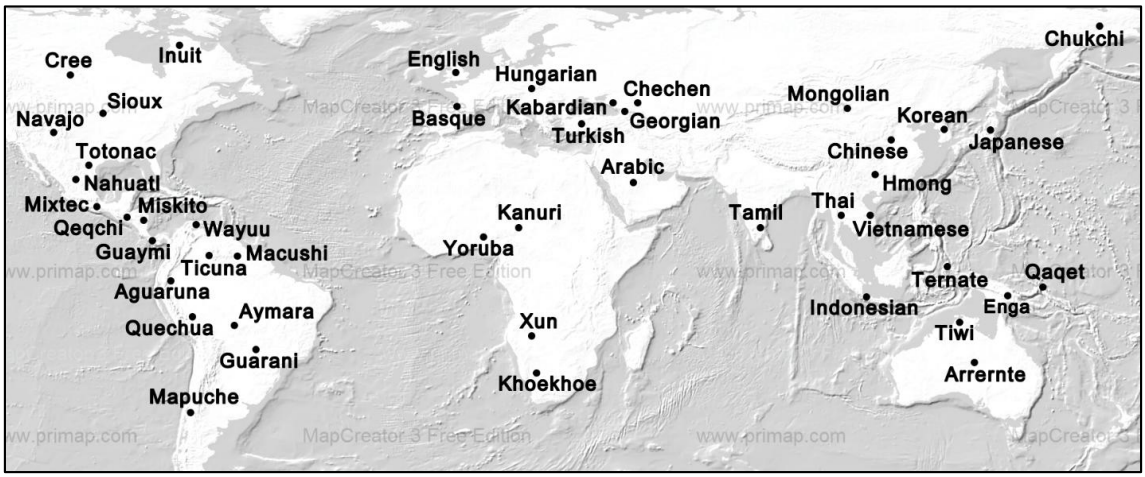
[13] This is because, in a sample of 81 observations, correlation coefficients are statistically significant at a 1% probability level if their absolute values are greater than 0.285.

included in the extended sample, plus 9 ones from the original sample (Arabic, Chinese, Filipino, Finnish, Fula, Malagasy, Nahuatl, Quechua and Vietnamese). These last observations have been selected to match the phylogenetic and geographic distribution of the original sample. All the languages of sub-sample 1 (both original and additional) are represented in figure 7.

Figure 7: Languages from sub-sample 1

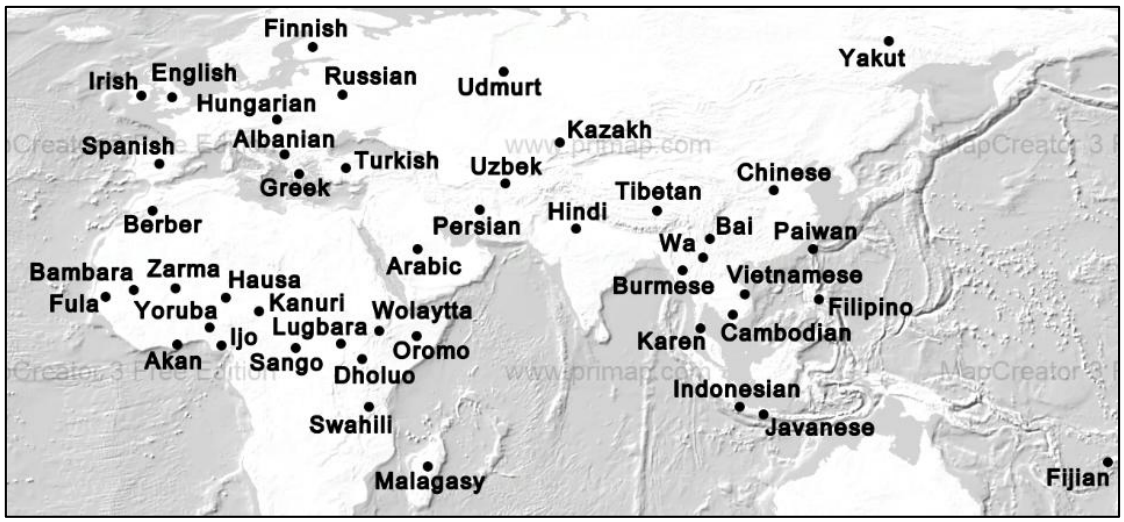
The second sub-sample, in turn, is formed by one observation from each family of the extended sample, and it includes the following 45 languages: Aguaruna, Arabic, Arrernte, Aymara, Basque, Chechen, Chinese, Chukchi, Cree, Enga, English, Georgian, Guarani, Guaymi, Hmong, Hungarian, Indonesian, Inuit, Japanese, Kabardian, Kanuri, Khoekhoe, Korean, Macushi, Mapuche, Miskito, Mixtec, Mongolian, Nahuatl, Navajo, Qaqet, Qeqchi, Quechua, Sioux, Tamil, Ternate, Thai, Ticuna, Tiwi, Totonac, Turkish, Vietnamese, Wayuu, Xun and Yoruba. All those languages are depicted in figure 8.

Figure 8: Languages from sub-sample 2



The third sub-sample, conversely, uses all the observations that belong to the nine families that have more than one language in the sample (see figure 9). It also includes 45 languages, which are Spanish, Irish, English, Russian, Albanian, Greek, Persian and Hindi (Indo-European); Fula, Bambara, Akan, Yoruba, Ijo, Sango and Swahili (Niger-Congo); Malagasy, Paiwan, Filipino, Indonesian, Javanese and Fijian (Austronesian); Berber, Hausa, Wolaytta, Oromo and Arabic (Afro-Asiatic); Chinese, Tibetan, Bai, Burmese and Karen (Sino-Tibetan); Turkish, Uzbek, Kazakh and Yakut (Turkic); Zarma, Kanuri, Lugbara and Dholuo (Nilo-Saharan); Hungarian, Finnish and Udmurt (Uralic); and Wa, Vietnamese and Cambodian (Austro-Asiatic).

Figure 9: Languages from sub-sample 3

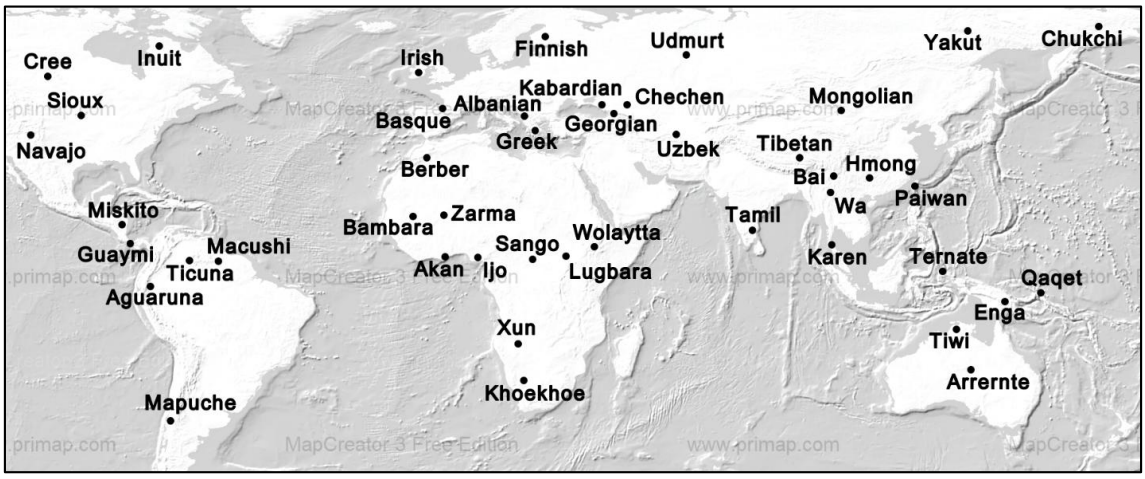


In order to avoid possible biases related to the high proportion of very large languages, our fourth sub-sample encompasses the five smallest languages from each of the nine regions defined in the original sample (see figure 10). The observations in this sub-sample correspond to Inuit, Cree, Sioux, Navajo and Miskito (North America); Guaymi, Ticuna, Macushi, Aguaruna and Mapuche (South America); Berber, Bambara, Zarma, Akan and Ijo (West Africa); Wolaytta, Lugbara, Sango, Xun and Khoekhoe (East Africa); Irish, Basque, Albanian, Greek and Finnish (Europe); Kabardian, Georgian, Chechen, Uzbek and Tamil (West Asia); Udmurt, Tibetan, Mongolian, Yakut and Chukchi (North Asia); Paiwan, Hmong, Bai, Wa and Karen (South Asia); and Ternate, Tiwi, Arrernte, Enga and Qaqet (Australasia).

With these four sub-samples, we calculated standard correlation coefficients for the typological complexity measures (*Phoncomp* and *Morphcomp*) and the corpus-based

language ratios (*Phonword* and *Wordclaus*). Later on, we applied the structure-conduct-performance approach in order to recalculate those coefficients, so that the new SCP coefficients could be compared with the standard ones (and with the correlation coefficients from the original sample).

Figure 10: Languages from sub-sample 4



The regression equations used to calculate the SCP coefficients for sub-sample 1 were exactly the same ones that we used for the original sample. However, when running equations 3 and 4 for sub-sample 2, we were not able to use the phylogenetic characteristics of languages as independent variables, since each language belongs to a different family. In the case of sub-sample 3, we did not use the variables *Northamerica* and *Southamerica*, because in that sub-sample there are no languages from the American continent. Finally, when dealing with sub-sample 4, we did not use the variables *Austronesian* and *Austroasiatic*, since they correspond to families with only one language in that sub-sample. We could not use the variable *Major*, either, since in sub-sample 4 there are no major languages.[14]

The main results of our analysis are summarized in table 4. We see that all the estimated coefficients have the same signs (either positive or negative) in the four sub-samples, and these signs are the same ones that we found previously for the original sample and the extended sample. Another general result is that the SCP correlation coefficients display larger absolute values than the corresponding standard coefficients (just as it happens in the other samples). Moreover, it holds that all the SCP coefficients

[14] All this can be seen in the tables of appendix 5, which report the corresponding regression equation coefficients.

are significant at a 1% probability level.

Table 4: Comparison between sub-samples

Concept	Sub-sample 1	Sub-sample 2	Sub-sample 3	Sub-sample 4
Standard correlation				
Phonword vs. Wordclaus	-0.7236	-0.6783	-0.7883	-0.7212
Phoncomp vs. Phonword	-0.6046	-0.5488	-0.4968	-0.5842
Morphcomp vs. Phonword	0.5636	0.4497	0.6282	0.5294
Phoncomp vs. Wordclaus	0.4614	0.5569	0.4010	0.5277
Morphcomp vs. Wordclaus	-0.3243	-0.3152	-0.5120	-0.4272
Phoncomp vs. Morphcomp	-0.2919	-0.2386	-0.4003	-0.3931
SCP correlation				
Phonword vs. Wordclaus	-0.9507	-0.9430	-0.9690	-0.9331
Phoncomp vs. Phonword	-0.6746	-0.6363	-0.5984	-0.7267
Morphcomp vs. Phonword	0.6514	0.5693	0.7548	0.6777
Phoncomp vs. Wordclaus	0.6280	0.6006	0.5415	0.6577
Morphcomp vs. Wordclaus	-0.5180	-0.4825	-0.6631	-0.6106
Phoncomp vs. Morphcomp	-0.4405	-0.4051	-0.4336	-0.5929

Table 4 also shows that one standard correlation coefficient (*Phoncomp* vs. *Morphcomp)* fails to be statistically significant at a 5% probability level in sub-samples 1 and 2. When applying the SCP approach, however, that coefficient increases its negative value from 0.2919 to 0.4405 (sub-sample 1) and from 0.2386 to 0.4051 (sub-sample 2), and therefore it becomes significant at a 1% probability level.

In sub-samples 1 and 2 there is another standard correlation coefficient (*Morphcomp* vs. *Wordclaus*) that is significant at a 5% probability level but not at a 1% probability level. However, when we apply the structure-conduct-performance approach, this coefficient increases its negative value from 0.3243 to 0.5180 (in sub-sample 1) and from 0.3152 to 0.4825 (in sub-sample 2), and those changes make it statistically significant at a 1% probability level.

The results that remain robust for all the SCP estimations, under the original sample, the extended sample and the four sub-samples, are therefore the following:

a) *Phonword* and *Wordclaus* are negatively correlated, and their SCP coefficients are always statistically significant at a 1% probability level.

b) *Phoncomp* and *Morphcomp* are also negatively correlated, and their SCP coefficients are significant at a 1% probability level (although, in some cases, the corresponding standard correlation coefficients are not significant).

c) The alternative measure of the language complexity trade-off between phonological and morphological complexity (*Phoncomp* vs. *Phonword*) also displays SCP coefficients

that are significant at a 1% probability level.

d) Finally, the alternative measure of the trade-off between morphological and syntactic complexity (*Morphcomp* vs. *Wordclaus*) displays SCP coefficients that are always significant at a 1% probability level (although, in some cases, the corresponding standard correlation coefficients are not significant at that level).

7. Concluding remarks

The results reported in the previous sections give support to the idea that, for an analysis of the possible existence of language complexity trade-offs, it is helpful to use an approach that links the relevant variables through a process. In our case, we postulate that the language outcomes that we observe empirically (which can be measured using different metrics applied to actual texts) are in some way determined by certain typological variables, related to the grammar of the different languages. Those variables, in turn, may be influenced by some characteristics of the languages, derived from phylogenetic, geographic, demographic and sociological phenomena.

One feasible way to implement this idea is to use the so-called "structure-conduct-performance paradigm", that we have taken from the literature about industrial economics. For our problem, we identified some phylogenetic, geographic, demographic and sociological characteristics of languages as structural variables, their typological characteristics as conduct variables, and some corpus-based language ratios as performance variables.

With that framework, we conducted a correlation and regression analysis that initially consisted of finding the influence of the structural variables on two typological measures of complexity. After that, we looked for the effect of the typological variables on two empirical variables (language ratios). Using those series, we finally found the corresponding correlation coefficients for the typological complexity measures, for the empirical measures, and also for the typological measures against the empirical ones.

In order to do all that, we previously ran a set of least-square regressions, in order to replace some variables by the fitted values of those regressions (which function as "instrumental variables"). Here this occurred with the typological complexity measures (that were replaced by the outcomes of regressions against the structural variables) and with the language ratios (that were replaced by the outcomes of regressions against the underlying typological variables).

As a consequence of this procedure, we found a relatively high negative correlation, indicative of a possible trade-off, between phonological complexity and morphological complexity (which is something that is not clear if we only use standard correlation coefficients). We also found that the negative correlation between phonemes per word and words per clause, and the negative correlation between morphological complexity and words per clause, which indicate a trade-off between morphology and syntax, increased when we applied the SCP model. And the same occurs with the correlation between phonological complexity and phonemes per word, which can be seen as an alternative measure of the trade-off between phonology and morphology.

Later on, we tested our model using an extended sample that encompasses the 45 languages of our original sample, plus 36 additional languages. With that extended sample we also built four sub-samples of 45 languages each. The first of those sub-samples comprises the additional languages, plus nine languages from the original sample; the second one consists of languages from different families; the third sub-sample has all the observations from the nine main language families; and the fourth sub-sample encompasses the five smallest languages from each of the nine world regions. The basic results remained the same, in the sense that the signs of the SCP correlation coefficients did not change, and all of them increased in absolute value (when compared to the corresponding standard correlation coefficients).

All the results reported in this paper, however, can be considered preliminary, since they were obtained using a short and relatively peculiar text, a relatively small sample of languages, and few and not very sophisticated complexity measures. Hopefully, however, the method outlined here may be useful for researchers that have access to other databases, opening a path to detecting the functioning of the language system within a more integrated context. We also believe that we have contributed to reconcile some opposing results that appear in the literature about language complexity trade-offs, by relating the values of our typological and corpus-based variables as part of a sequential process.

References

Angrist, Joshua & Jörn Pischke (2009) *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Benítez, Antonio, Sihan Chen & David Gil (2024) The Absence of a Trade-Off between Morphological and Syntactic Complexity. *Frontiers in Language Sciences* 3: 1340493.

Bentz, Chistian, Ximena Gutiérrez, Olga Sozinova & Tanja Samardzic (2023) Complexity Trade-Offs and Equi-Complexity in Natural Languages: A Meta-Analysis. *Linguistics Vanguard* 9: 9-25.

Bonamente, Massimiliano (2022) *Statistics and Analysis of Scientific Data*, 3rd edition. New York, Springer.

Chen, Sihan, David Gil, Sergey Gaponov, Jana Reifegerste, Tessa Yuditha, Tatiana Tatarinova, Ljiljana Progovac & Antonio Benítez (2024) Linguistic Correlates of Societal Variation: A Quantitative Analysis. *PLOS One* 19(4): e0300838.

Coloma, Germán (2016) An Optimization Model of Global Language Complexity. *Glottometrics* 35: 49-63.

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

Coloma, Germán (2017b) Complexity Trade-Offs in the 100-Language WALS Sample. *Language Sciences* 59: 148-158.

Coloma, Germán (2022) Correlation between Linguistic Measures: An Extended Analysis. *Studies in Linguistics and Literature* 6(4): 109-132.

Coloma, Germán (2024) An Extended Synergetic Model of Language Phonology. *Linguistic Exploration* 1: 12-35.

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

Fenk-Oczlon, Gertraud & August Fenk (2008) Complexity Trade-Offs Between the Subsystems of Language. In M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change*, 43-65. Amsterdam: John Benjamins.

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

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

Kumar, Neeraj & Pooja Choudhary (2024) Structure, Conduct and Performance in Industrial Organization: A Contemporary Examination of Causal Relationships. *Empirical Economics Letters* 23(2): 33-52.

Levshina, Natalia (2020) Efficient Trade-Offs as Explanations in Functional Linguistics: Some Problems and an Alternative Proposal. *Revista da Abralin* 19(3): 50-78.

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 & François Pellegrino (2023) Towards Robust Complexity Indices in Linguistic Typology. *Studies in Language* 47(4): 789-829.

Perloff, Jeffrey, Larry Karp & Amos Golan (2007) *Estimating Market Power and Strategies*. New York: Cambridge University Press.

Shcherbakova, Olena, Volker Gast, Damián Blasi, Hedvig Skirgard, Russell Gray & Simon Greenhill (2023) A Quantitative Global Test of the Complexity Trade-Off Hypothesis: The Case of Nominal and Verbal Grammatical Marking. *Linguistics Vanguard* 9: 155-167.

Shosted, Ryan (2006) Correlating Complexity: A Typological Approach. *Linguistic Typology* 10: 1-40.

Appendix 1: Languages included in the original sample

Language	Genus	Family	Region	Size	Use
Arabic	Semitic	Afro-Asiatic	West Asia	Major	Exoteric
Arrernte	Arandic	Pama-Nyungan	Australasia	Small	Local
Aymara	Aymaran	Aymaran	South America	Medium	Local
Bambara	Mande	Niger-Congo	West Africa	Medium	Exoteric
Burmese	Burmic	Sino-Tibetan	South Asia	Large	Local
Cambodian	Khmer	Austro-Asiatic	South Asia	Large	Local
Chinese	Sinitic	Sino-Tibetan	North Asia	Major	Exoteric
Dholuo	Nilotic	Nilo-Saharan	East Africa	Medium	Local
Enga	Engan	Trans-New Guinea	Australasia	Small	Local
English	Germanic	Indo-European	Europe	Major	Exoteric
Fijian	Oceanic	Austronesian	Australasia	Small	Local
Filipino	Philippine	Austronesian	Australasia	Large	Exoteric
Finnish	Finnic	Uralic	Europe	Medium	Local
Fula	Atlantic	Niger-Congo	West Africa	Large	Exoteric
Georgian	Kartvelian	South Caucasian	West Asia	Medium	Local
Guarani	Guaranitic	Tupian	South America	Medium	Local
Hausa	Chadic	Afro-Asiatic	West Africa	Large	Exoteric
Hindi	Indic	Indo-European	West Asia	Major	Exoteric
Hmong	Hmongic	Hmong-Mien	South Asia	Medium	Local
Hungarian	Ugric	Uralic	Europe	Large	Local
Indonesian	Malayic	Austronesian	Australasia	Major	Exoteric
Inuit	Eskimo	Eskimo-Aleut	North America	Small	Local
Japanese	Japonic	Japonic	North Asia	Major	Local
Kanuri	Saharan	Nilo-Saharan	West Africa	Large	Exoteric
Kazakh	Kipchak	Turkic	North Asia	Large	Local
Khoekhoe	Khoe	Khoe-Kwadi	East Africa	Small	Local
Korean	Koreanic	Koreanic	North Asia	Large	Local
Malagasy	Malagasy	Austronesian	East Africa	Large	Local
Mapuche	Araucanian	Araucanian	South America	Small	Local
Mixtec	Mixtecan	Oto-Manguean	North America	Small	Local
Nahuatl	Aztecan	Uto-Aztecan	North America	Medium	Local
Navajo	Athabaskan	Na-Dene	North America	Small	Local
Oromo	Cushitic	Afro-Asiatic	East Africa	Large	Local
Qeqchi	Quichean	Mayan	North America	Medium	Local
Quechua	Quechuan	Quechuan	South America	Medium	Exoteric
Russian	Slavic	Indo-European	Europe	Major	Exoteric
Spanish	Romance	Indo-European	Europe	Major	Exoteric
Swahili	Bantoid	Niger-Congo	East Africa	Large	Exoteric
Tamil	Tamil-Kannada	Dravidian	West Asia	Large	Local
Thai	Kam-Tai	Tai-Kadai	South Asia	Large	Exoteric
Tibetan	Tibetic	Sino-Tibetan	North Asia	Medium	Local
Turkish	Oghuz	Turkic	West Asia	Large	Local
Vietnamese	Vietic	Austro-Asiatic	South Asia	Large	Local
Wayuu	Guajiro	Arawakan	South America	Small	Local
Yoruba	Volta-Niger	Niger-Congo	West Africa	Large	Local

Appendix 2: Typological (conduct) variables

Language	Consonants	Vowels	Tones	Cases	Genders	Inflections
Arabic	29	6	1	1	2	6
Arrernte	27	4	1	8	1	4
Aymara	27	6	1	6	1	8
Bambara	20	18	2	1	1	2
Burmese	34	9	4	8	1	2
Cambodian	16	20	1	1	1	1
Chinese	19	6	4	1	1	1
Dholuo	25	9	3	1	1	6
Enga	16	5	5	5	1	6
English	24	11	1	2	1	2
Fijian	16	5	1	1	1	6
Filipino	16	5	1	1	2	2
Finnish	13	8	1	10	1	2
Fula	27	10	1	1	8	2
Georgian	28	5	1	6	1	8
Guarani	19	12	1	1	1	4
Hausa	28	10	2	1	2	6
Hindi	34	11	1	2	2	2
Hmong	58	9	7	1	1	2
Hungarian	25	14	1	10	1	4
Indonesian	18	6	1	1	1	4
Inuit	14	6	1	8	1	4
Japanese	16	5	2	1	1	4
Kanuri	25	7	2	6	1	8
Kazakh	20	11	1	6	1	6
Khoekhoe	31	8	3	2	3	6
Korean	19	18	1	6	1	6
Malagasy	29	4	1	1	1	4
Mapuche	22	6	1	2	1	8
Mixtec	24	8	5	1	5	4
Nahuatl	20	5	3	1	1	8
Navajo	28	16	2	1	1	5
Oromo	24	10	2	6	2	6
Qeqchi	26	10	1	1	6	4
Quechua	25	3	1	8	1	8
Russian	36	6	1	6	3	4
Spanish	19	5	1	1	2	4
Swahili	32	5	1	1	5	4
Tamil	15	10	1	6	3	2
Thai	21	9	5	1	1	2
Tibetan	28	8	2	1	1	4
Turkish	22	8	1	6	1	6
Vietnamese	22	11	8	1	1	1
Wayuu	14	12	1	1	2	6
Yoruba	18	11	3	1	1	6
Average	**23.76**	**8.69**	**2.02**	**3.20**	**1.73**	**4.44**

Appendix 3: Empirical variables and complexity measures

Language	Clauses	Words	Phonemes	Phonword	Wordclaus	Phoncomp	Morphcomp
Arabic	9	85	488	5.74	9.44	35	9
Arrernte	12	73	436	5.97	6.08	31	13
Aymara	9	80	511	6.39	8.89	33	15
Bambara	9	129	391	3.03	14.33	56	4
Burmese	5	42	300	7.14	8.40	70	11
Cambodian	9	122	549	4.50	13.56	36	3
Chinese	10	98	421	4.30	9.80	43	3
Dholuo	10	116	454	3.91	11.60	52	8
Enga	10	96	341	3.55	9.60	41	12
English	9	113	383	3.39	12.56	35	5
Fijian	9	140	556	3.97	15.56	21	8
Filipino	9	116	546	4.71	12.89	21	5
Finnish	9	76	441	5.80	8.44	21	13
Fula	6	91	358	3.93	15.17	37	11
Georgian	9	70	418	5.97	7.78	33	15
Guarani	9	81	455	5.62	9.00	31	6
Hausa	12	166	648	3.90	13.83	48	9
Hindi	8	125	467	3.74	15.63	45	6
Hmong	9	149	464	3.11	16.56	121	4
Hungarian	10	100	431	4.31	10.00	39	15
Indonesian	9	108	594	5.50	12.00	24	6
Inuit	9	61	607	9.95	6.78	20	13
Japanese	9	89	444	4.99	9.89	26	6
Kanuri	9	62	404	6.52	6.89	39	15
Kazakh	8	90	575	6.39	11.25	31	13
Khoekhoe	8	78	299	3.83	9.75	55	11
Korean	7	60	381	6.35	8.57	37	13
Malagasy	10	126	564	4.48	12.60	33	6
Mapuche	9	75	360	4.80	8.33	28	11
Mixtec	7	118	394	3.34	16.86	64	10
Nahuatl	9	85	468	5.51	9.44	35	10
Navajo	8	90	387	4.30	11.25	60	7
Oromo	8	84	440	5.24	10.50	44	14
Qeqchi	9	133	530	3.98	14.78	36	11
Quechua	9	77	581	7.55	8.56	28	17
Russian	9	97	468	4.82	10.78	42	13
Spanish	9	97	425	4.38	10.78	24	7
Swahili	8	72	367	5.10	9.00	37	10
Tamil	9	79	541	6.85	8.78	25	11
Thai	11	131	480	3.66	11.91	66	4
Tibetan	9	81	446	5.51	9.00	44	6
Turkish	9	66	431	6.53	7.33	30	13
Vietnamese	7	117	334	2.85	16.71	110	3
Wayuu	9	56	301	5.38	6.22	26	9
Yoruba	8	120	373	3.11	15.00	51	8
Average	**8.84**	**96.00**	**450.1**	**4.98**	**10.93**	**41.42**	**9.38**

Appendix 4: Additional languages included in the extended sample

Language	Genus	Family	Region	Size	Use
Aguaruna	Jivaroan	Jivaroan	South America	Small	Local
Akan	Kwa	Niger-Congo	West Africa	Medium	Local
Albanian	Albanic	Indo-European	Europe	Medium	Local
Bai	Macro-Bai	Sino-Tibetan	South Asia	Medium	Local
Basque	Vasconic	Vasconic	Europe	Small	Local
Berber	Berberic	Afro-Asiatic	West Africa	Medium	Local
Chechen	Nakh	Northeast Caucasian	West Asia	Medium	Local
Chukchi	Chukotkan	Paleo-Siberian	North Asia	Small	Local
Cree	Algonquian	Algic	North America	Small	Local
Greek	Hellenic	Indo-European	Europe	Large	Exoteric
Guaymi	Guaymic	Chibchan	South America	Small	Local
Ijo	Ijoid	Niger-Congo	West Africa	Medium	Local
Irish	Celtic	Indo-European	Europe	Small	Local
Javanese	Javanese	Austronesian	Australasia	Large	Local
Kabardian	Circassian	Northwest Caucasian	West Asia	Medium	Local
Karen	Karenic	Sino-Tibetan	South Asia	Medium	Local
Lugbara	Moru-Madi	Nilo-Saharan	East Africa	Medium	Local
Macushi	Pemong	Cariban	South America	Small	Local
Miskito	Misumalpan	Misumalpan	North America	Small	Local
Mongolian	Mongolic	Mongolic	North Asia	Medium	Local
Paiwan	South Formosan	Austronesian	South Asia	Small	Local
Persian	Iranian	Indo-European	West Asia	Large	Exoteric
Qaqet	Baining	East Papuan	Australasia	Small	Local
Sango	Ubangi	Niger-Congo	East Africa	Medium	Exoteric
Sioux	Dakotan	Siouan	North America	Small	Local
Ternate	Halmaheran	West Papuan	Australasia	Small	Local
Ticuna	Ticunan	Ticunan	South America	Small	Local
Tiwi	Tiwian	Tiwian	Australasia	Small	Local
Totonac	Totonacan	Totonacan	North America	Small	Local
Udmurt	Permic	Uralic	North Asia	Small	Local
Uzbek	Karluk	Turkic	West Asia	Large	Local
Wa	Palaungic	Austro-Asiatic	South Asia	Medium	Local
Wolaytta	Omotic	Afro-Asiatic	East Africa	Medium	Local
Xun	Ju-Kung	Kxa	East Africa	Small	Local
Yakut	Siberian	Turkic	North Asia	Small	Local
Zarma	Songhay	Nilo-Saharan	West Africa	Medium	Local

Language	Consonants	Vowels	Tones	Cases	Genders	Inflections
Aguaruna	15	8	1	6	1	6
Akan	27	10	3	1	1	6
Albanian	29	7	1	4	3	7
Bai	21	15	5	1	1	2
Basque	23	5	1	10	1	4
Berber	32	3	1	2	2	6
Chechen	58	7	1	10	5	10
Chukchi	15	7	1	9	1	4
Cree	10	7	1	1	2	6
Greek	18	5	1	3	3	4
Guaymi	25	16	1	1	1	3
Ijo	20	18	2	1	1	6
Irish	35	11	1	2	2	2
Javanese	21	6	1	1	1	4
Kabardian	45	2	1	4	1	8
Karen	28	9	6	1	1	2

Language	Consonants	Vowels	Tones	Cases	Genders	Inflections
Lugbara	36	7	3	1	1	4
Macushi	10	12	1	5	2	4
Miskito	14	6	1	1	1	4
Mongolian	17	14	1	8	1	2
Paiwan	23	4	1	1	1	4
Persian	23	6	1	2	1	4
Qaqet	16	4	1	1	8	3
Sango	26	12	3	1	1	1
Sioux	31	8	1	1	1	10
Ternate	19	5	1	1	2	2
Ticuna	11	6	10	1	5	4
Tiwi	17	4	1	1	2	4
Totonac	17	6	1	1	1	4
Udmurt	26	7	1	10	1	3
Uzbek	25	7	1	6	1	6
Wa	33	9	2	1	1	1
Wolaytta	29	10	2	3	2	3
Xun	94	5	4	1	4	2
Yakut	20	16	1	8	1	7
Zarma	20	16	4	1	1	2
Average	**25.81**	**8.33**	**1.92**	**3.11**	**1.81**	**4.28**

Language	Clauses	Words	Phonemes	Phonword	Wordclaus	Phoncomp	Morphcomp
Aguaruna	6	68	413	6.07	11.33	23	13
Akan	8	112	381	3.40	14.00	57	8
Albanian	7	114	402	3.53	16.29	36	14
Bai	8	115	376	3.27	14.38	96	4
Basque	7	83	401	4.83	11.86	28	15
Berber	9	76	306	4.03	8.44	35	10
Chechen	7	87	408	4.69	12.43	65	25
Chukchi	7	62	335	5.40	8.86	22	14
Cree	8	52	382	7.35	6.50	17	9
Greek	9	104	479	4.61	11.56	23	10
Guaymi	7	70	265	3.79	10.00	41	5
Ijo	12	175	625	3.57	14.58	56	8
Irish	8	129	406	3.15	16.13	46	6
Javanese	8	84	416	4.95	10.50	27	6
Kabardian	9	56	353	6.30	6.22	47	13
Karen	10	151	323	2.14	15.10	82	4
Lugbara	9	115	398	3.46	12.78	57	6
Macushi	9	81	323	3.99	9.00	22	11
Miskito	7	87	361	4.15	12.43	20	6
Mongolian	7	81	392	4.84	11.57	31	11
Paiwan	9	93	367	3.95	10.33	27	6
Persian	9	91	483	5.31	10.11	29	7
Qaqet	17	158	774	4.90	9.29	20	12
Sango	9	142	443	3.12	15.78	62	3
Sioux	9	65	346	5.32	7.22	39	12
Ternate	8	77	296	3.84	9.63	24	5
Ticuna	11	101	495	4.90	9.18	71	10
Tiwi	8	70	428	6.11	8.75	21	7
Totonac	9	78	357	4.58	8.67	23	6
Udmurt	8	68	383	5.63	8.50	33	14
Uzbek	8	87	520	5.98	10.88	32	13
Wa	11	137	450	3.28	12.45	51	3
Wolaytta	6	63	316	5.02	10.50	49	8
Xun	9	95	277	2.92	10.56	114	7

Language	Clauses	Words	Phonemes	Phonword	Wordclaus	Phoncomp	Morphcomp
Yakut	9	71	428	6.03	7.89	36	16
Zarma	8	94	292	3.11	11.75	84	4
Average	**8.61**	**94.22**	**400.0**	**4.49**	**10.98**	**42.94**	**9.19**

Appendix 5: Results of the regressions to generate the instrumental variables

Table 5a: Estimated coefficients for the conduct variable regression equations

Concept / Equation	Original sample		Extended sample	
	Phoncomp	Morphcomp	Phoncomp	Morphcomp
Europe	26.62378	13.06854	30.52621	15.07487
Westafrica	58.34761	9.13256	71.13582	10.18922
Eastafrica	53.46027	10.02193	70.87746	9.21126
Northamerica	43.20438	10.02185	33.79273	9.30158
Southamerica	30.02958	11.38458	32.36929	10.69831
Westasia	37.95711	10.73521	39.37347	13.80060
Southasia	96.45567	4.32266	77.18668	7.14329
Northasia	45.22189	7.47970	30.78381	11.80711
Australasia	36.37093	12.65712	35.68510	9.65995
Indoeuro	16.07338	-3.07230	0.52640	-4.94485
Afroasiatic	5.27552	-0.73613	-23.72792	-0.26315
Nigercongo	0.31799	-3.36223	-20.01028	-2.99626
Austronesian	-6.24538	-6.82799	-21.88500	-2.84199
Sinotibetan	-2.29499	-0.39494	4.94192	-2.99708
Turkic	2.38352	0.77225	-0.67715	0.47426
Uralic	10.36821	-0.85137	-0.30011	-0.24297
Austroasiatic	-9.98265	-4.44295	-11.25183	-4.61037
Nilosaharan	-2.10444	0.26941	-15.20858	-1.78633
Major	-6.41548	-1.40394	8.35493	-3.40300
Large	-13.47302	3.12029	-2.86865	0.62918
Medium	-0.51096	0.44537	4.93273	0.14288
Exoteric	-2.61502	-0.25898	-3.12176	0.28653

Table 5b: Estimated coefficients for the performance variable regression equations

Concept / Equation	Original sample		Extended sample	
	Phonword	Wordclaus	Phonword	Wordclaus
Constant	5.64371	9.67677	5.31070	9.43462
Consonants	-0.01723	0.04535	-0.02841	0.03824
Vowels	-0.06861	0.14082	-0.07335	0.19692
Tones	-0.24139	0.41040	-0.13643	0.23769
Cases	0.24703	-0.41043	0.18695	-0.24574
Genders	-0.08643	0.37208	-0.02387	0.20017
Inflections	0.04155	-0.27096	0.11321	-0.26249

Table 5c: Estimated coefficients for the conduct equations using sub-samples

Concept	Sub-sample 1	Sub-sample 2	Sub-sample 3	Sub-sample 4
Phoncomp equation				
Europe	33.04413	37.04032	-25.91416	18.01316
Westafrica	71.10913	58.11879	2.22274	86.57495
Eastafrica	78.91920	84.50000	1.33583	84.50000
Northamerica	26.64875	33.71783		31.20000
Southamerica	38.84544	33.47715		37.00000
Westasia	48.55207	39.06309	-19.19686	35.59680
Southasia	63.29821	105.98927	7.49705	59.55448
Northasia	30.66960	30.36029	-25.60281	23.03440
Australasia	33.36027	27.34233		27.40000
Indoeuro	-0.27781		62.83778	16.10180
Afroasiatic	-28.23516		47.96036	-56.53239
Nigercongo	-16.46679		54.97021	-42.68229
Austronesian	-25.22767		27.27679	
Sinotibetan	19.78839		73.91213	13.62396
Turkic	-7.27020		57.18968	11.23546
Uralic	-5.23500		58.04168	-0.02124
Austroasiatic	15.16429		60.52974	
Nilosaharan	-5.27043		59.24715	-28.03239
Major	13.59602	11.04970	6.25957	
Large	3.31873	-6.06694	-3.38646	-13.10211
Medium	0.75627	5.26977	-0.30745	12.99492
Exoteric	-9.98347	-14.10371	-7.79691	2.76231
Morphcomp equation				
Europe	15.12728	14.16131	0.40854	13.90954
Westafrica	10.72446	10.92877	-1.68595	10.39630
Eastafrica	7.64925	9.00000	-2.17946	9.00000
Northamerica	8.10355	8.97782		9.40000
Southamerica	10.49544	9.88215		10.00000
Westasia	15.25222	14.95730	-2.39127	16.92625
Southasia	9.33541	2.75257	0.26109	5.61340
Northasia	13.70186	12.51397	-0.14231	12.40608
Australasia	7.78358	10.23743		9.80000
Indoeuro	-7.38708		9.70140	-2.15383
Afroasiatic	-2.57875		12.58079	-0.37561
Nigercongo	-5.00649		10.16647	-3.66820
Austronesian	-3.15467		6.16275	
Sinotibetan	-6.66691		6.44363	-2.88842
Turkic	-0.31476		13.82215	1.99965
Uralic	-2.15571		13.76593	0.50346
Austroasiatic	-7.91426		2.19881	
Nilosaharan	-6.66912		11.26440	-4.37561
Major	-4.89476	-9.68577	-0.38999	
Large	0.67543	-1.05937	1.59286	-4.33163
Medium	2.48227	1.59983	-1.56543	-0.32254
Exoteric	1.04055	3.26122	-1.22337	-0.61297

Table 5d: Estimated coefficients for the performance equations using sub-samples

Concept	Sub-sample 1	Sub-sample 2	Sub-sample 3	Sub-sample 4
Phonword equation				
Constant	5.14520	5.89593	5.10663	5.41936
Consonants	-0.03103	-0.02785	-0.02363	-0.02980
Vowels	-0.09934	-0.09853	-0.07160	-0.08568
Tones	-0.08122	-0.17109	-0.15133	-0.12892
Cases	0.11643	0.12459	0.22456	0.16185
Genders	0.00246	-0.04772	-0.02618	0.01501
Inflections	0.19458	0.12973	0.07261	0.09411
Wordclaus equation				
Constant	9.28790	8.00327	10.17406	8.41738
Consonants	0.03404	0.03153	0.01678	0.04728
Vowels	0.23735	0.18419	0.21831	0.25203
Tones	0.15602	0.43856	0.24184	0.22274
Cases	-0.10254	-0.04514	-0.46596	-0.13239
Genders	0.12565	0.20088	0.31688	-0.02080
Inflections	-0.28346	-0.27171	-0.08900	-0.23713

Appendix 6: Examples of the text of "The North Wind and the Sun"

English (Europe, Indo-European)

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.

Arabic (West Asia, Afro-Asiatic)

Kaanat riihu al-shamaali tatazhaadalu wa al-shamsu fii ayyin minhumaa kaanat aqwaa min al-ukraa, wa id bi-musaafirin yatla'u mutalaffi'an bi-'abaa'atin sami ikatin. Fa ittafaqataa 'alaa i'tibaari al-saabiqi fii izhbaari al-musaafiri 'alaa kal'i 'abaa'atihi al-aqwaa. 'Asafat riihu al-shamaali bi-aqsaa maa istataa'at min quuwatin. Wa laakin kullumaa izdaada al'asfu, izdaada al-musaafiru tadatturan bi-'abaa'atihi, ilaa an usqita fii yadi al-riih fatakallat 'an muhaawalatihaa. Ba'da'idin sata'ati al-shamsu bi-dif'ihaa, famaakaana min al-musaafiri illaa an kala'a 'abaa'atahu 'alaa al-tauu. Wa hakadaa idtarrat riihu al-shamaali ilaa al-i'tiraafi bi-anna al-shamsa kaanat hiya al-aqwaa.

Chinese (North Asia, Sino-Tibetan)

Yǒu yì huí běifēng gēn tàiyáng zhèngzàinar zhēnglùn shéide běnshi dà, shuōzhe shuōzhe láile yíge zǒudàorde, shēnshang chuānzhe yíjiàn hòu páozi. Tāmen liǎ jiù shāngliang hǎo le shuō, "shéi néng xiān jiào zhège zǒudàorde bǎ tāde páozi tuōle xiàlai a, jiù suàn shéide běnshi dà". Hǎo, běifēng jiù shǐqi dà jìn lá jǐnguā jǐnguā, kěshì tā guāde yuè lìhai, nèige rén bǎ páozi guǒde yuè jǐn; daà mòliǎor běifēng méile fázi, zhǐhǎo jiù suànle. Yìhuǐr tàiyáng jiù chūlái rèrērde yí shài, nèi zǒudàorde mǎshàng jiù bǎ páozi tuōle xiàlai. Suǒyǐ běifēng bù néng bù chéngrèn dàodǐ háishi tàiyáng bǐ tā běnshi dà.

Indonesian (Australasia, Austronesian)

Sang Angin Utara dan Sang Matahari sedang berdebat tentang siapa diantaramereka yang paling hebat, ketika melintaslah seorang pelancong yang terbungkus dengan jubah hangatnya. Mereka setuju jika Sang Angin Utara berhasil membuat si pelancong tersebut membuka jubahnya, maka dialah yang menjadi terhebat diantara mereka. Dan Sang Angin Utara pun bertiup sekuat mungkin, namun semakin kuat ia bertiup semakin erat pulalah si pelancong memeluk jubahnya, sehingga pada akhirnya Sang Angin Utara itu menyerahlah. Sekarang tibalah giliran Sang Matahari untuk bersinar dengan hangatnya, dan saat itupun si pelancong membuka jubahnya sehingga membuat Sang Angin Utara harus mengakui bahwa Sang Mataharilah yang lebih hebat dari pada Sang Angin Utara itu sendiri.

Quechua (South America, Quechuan)

Huk p'unchaymi karu llaqtapi wichay manta kaq wayra intiwan rimasiarqanku, mayqinninkuchus sinchi kallpayupuni kasqankumanta. Chayllamanmi huk punchuyuq runa pasasiarqan rimasqankuq qayllanta. Chaymi wayraqa intita nirqan. Mayqinninkuchus haqay runata punchunta ch'ustirachisuman chayqa, chaypuni ancha kallpayuq hina riqsisqa kanqa nispa. Hinaspanmanmi wichaymanta kaq wayraqa sinchitapuni wayrayta qallarirqan, sinchita wayramuqtintaq runaqa qaqata punchunta hapirukurqan, wayraq thanimunan kama. Chayllamanmi intiñataq k'an niqtaraq kanch'arimurqan, hinan chay runaqa sinchita rupharispa punchunta ch'ustirukurqan. Chay p'unchaymi wichaymanta kaq wayraqa yacharqan, inti aswan kallpayuqpuni paymanta kasqanta.

Swahili (East Africa, Niger-Congo)

Upepo ulikuwa ukishindana na jua kuwa nani mwenye nguvu kupita mwenziwe. Mara akapita msafiri alichukua amevaa juba. Walipatana kuwa atakayemvua juba kwanza msafiri ndiyo mwenye nguvu. Upepo ukaanza kuvuma mwisho wa nguvu zake. Lakini kila ukizidi kuvuma ndiyo kwanza msafiri huzidi kulibanzia juba lake. Hata mwisho upepo ukakata tamaa. Jua likaanza kung'aa kwa ukali haikupata muda mara msafiri alivua juba lake. Na kwa hivyo upepo ukakiri kuwa jua lina nguvu kuliko yeye.

Vietnamese (South Asia, Austro-Asiatic)

Gió bâc và mat tròi cãi nhau xem ai manh hon, trong lúc dó mot du khách mac môt áo khoác âm di qua. Ho giao kèo vói nhau ràng ai là nguòi dâu tiên mà có thê bát nguòi du khách kia coi áo thì sẽ duoc coi là manh hon. Sau dó gió bâc bát dâu thôi manh hêt súc có thê, nhung càng thôi thì nguòi du khách càng giũ chat áo khoác và cuôi cùng gió bâc dã phai tù bo. Sau dó mat tròi suoi âm và nguòi du khách liên coi áo khoác. Kêt cuc là gió bâc phai thûa nhân ràng mat tròi là nguòi manh hon trong hai nguòi.