ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Coloma, Germán

Working Paper

A simultaneous-equation regression model of language complexity trade-offs

Serie Documentos de Trabajo, No. 597

Provided in Cooperation with:
University of CEMA, Buenos Aires

Suggested Citation: Coloma, Germán (2016) : A simultaneous-equation regression model of language complexity trade-offs, Serie Documentos de Trabajo, No. 597, ISBN 978-987-3940-08-8, Universidad del Centro de Estudios Macroeconómicos de Argentina (UCEMA), Buenos Aires

This Version is available at:
https://hdl.handle.net/10419/163255

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

UNIVERSIDAD DEL CEMA
Buenos Aires
Argentina

Serie
DOCUMENTOS DE TRABAJO

Área: Lingüística y Estadística

A SIMULTANEOUS-EQUATION REGRESSION MODEL OF LANGUAGE COMPLEXITY TRADE-OFFS

Germán Coloma

Octubre 2016
Nro. 597

ISBN 978-987-3940-08-8
Queda hecho el depósito que marca la Ley 11.723
Copyright – UNIVERSIDAD DEL CEMA

www.cema.edu.ar/publicaciones/doc_trabajo.html
UCEMA: Av. Córdoba 374, C1054AAP Buenos Aires, Argentina
ISSN 1668-4575 (impreso), ISSN 1668-4583 (en línea)
Editor: Jorge M. Streb; asistente editorial: Valeria Dowding <jae@cema.edu.ar>

Coloma, Germán
A simultaneous-equation regression model of language complexity trade-offs / Germán Coloma. - 1a ed . - Ciudad Autónoma de Buenos Aires : Universidad del CEMA, 2016.
21 p. ; 22 x 15 cm.

ISBN 978-987-3940-08-8

1. Ciencias Económicas. I. Título.
CDD 330

A Simultaneous-Equation Regression Model of Language Complexity Trade-Offs

Germán Coloma*

Abstract

In this paper we develop a statistical model of language complexity trade-offs using four typological measures (related to phonology, morphology, syntax and lexicon). The data come from the 100-language sample that appears in the World Atlas of Language Structures (WALS), and the trade-offs are calculated using different types of correlation coefficients. All those coefficients are statistically insignificant when they are computed using a standard (product-moment) methodology, but they become significant when we use simultaneous-equation regression methods, especially the ones based on seemingly unrelated regressions (SUR) and three-stage least squares (3SLS). These results are related to ideas suggested in the theoretical literature, especially in the one about language as a complex adaptive system.

Keywords: complexity trade-off, WALS, correlation, simultaneous-equation regression, complex adaptive system.

1. Introduction

A language complexity trade-off is a situation in which a higher level of complexity for a certain language component appears in correspondence to a lower level of complexity for another component. The literature about this topic can be divided between papers that show that languages usually exhibit complexity trade-offs, and papers that show that such trade-offs do not exist. Among the first group of papers we can cite contributions such as Nettle (1995) and Fenk-Oczlon and Fenk (2008), while in the second group there are articles like Shosted (2006) and Nichols (2009).

In general, the way in which the different authors assess the possible existence and significance of complexity trade-offs is some version of correlation analysis. Under that approach, two different measures of complexity (e.g., phonological and

* CEMA University; Av. Córdoba 374, Buenos Aires, C1054AAP, Argentina. Telephone: 54-11-6314-3000. E-mail: gcoloma@cema.edu.ar. I thank Gabriel Altmann, Damián Blasi, Guiomar Ciapuscio, Verónica Nercesián and Frans Plank for their comments to a previous version of this paper. The opinions expressed in this publication are my own, and are not necessarily the ones of CEMA University.

morphological complexity) are supposed to display a trade-off if they are negatively correlated between themselves, and the way to find that correlation (or its absence) is to calculate a coefficient based on the values of the different complexity measures in a sample of languages.[1]

If we look at the main methodological difference between the literature that finds statistically significant complexity trade-offs and the literature that does not find them, we see that one important point is the type of measures that they use. While in the first group authors generally rely on "empirical measures" (i.e., measures of complexity calculated using data from actual words or texts written in different languages), in the second group they generally use theoretical or "typological" measures (i.e., measures obtained from the grammars of the sample languages).

Another feature that we have found in previous work (Coloma 2014, 2016) is that language complexity trade-offs seem to be more important and statistically more significant if we measure them using partial correlation coefficients instead of standard correlation coefficients. This is related to the fact that, when we use a partial correlation coefficient, we are also including information from factors besides the two correlating variables. It is also linked to the idea that complexity variables can be determined by a system in which there are interactions among them, so each partial measure of complexity can be correlated to several other measures at the same time.

This reference to a system of relationships between the different complexity measures can be related to a branch of the theoretical literature that sees language as a self-organizing and self-regulating system whose properties come from the interaction of several constitutive, forming and control requirements. That literature is known as "synergetic linguistics", and its origins can be traced back to Köhler (1986, 1987). It is also related to another branch of the linguistic literature that sees language as a complex adaptive system (e.g., Beckner et al. 2009).

The aim of this paper is to look for the existence of trade-offs in a context where

[1] An alternative way to do that is to run a regression between the two variables under analysis. In that case, the relevant coefficient is the slope of the regression line obtained, which should also be negative if we are looking for evidence of a complexity trade-off.

language complexity is measured using theoretical variables (which is the one in which they have been harder to find), through a synergetic approach in which different factors interact. To do that we use a statistical methodology based on simultaneous-equation regressions, whose results allow to calculate different types of partial correlation coefficients between complexity measures. The analysis will be performed using the so-called "100-language sample" from the World Atlas of Language Structures (WALS), and complexity will be measured using binary variables that represent different concepts of phonological, morphological, syntactic and lexical complexity.

2. Description of the data

The WALS is a large database that compiles information about structural features from the grammars of the world's languages. In its current online version (Dryer and Haspelmath 2013), it contains data from 2679 languages and dialects, corresponding to 192 features that belong to different components of language structure.

The editors of the WALS have selected a sample of 100 languages for which they ask the authors of the different chapters of the atlas to include in their reports "if at all possible", and those languages are supposed to form a relatively balanced sample of genealogical and areal diversity.[2] Making use of the fact that we have more information about the languages that belong to this sample than the one available for the remaining languages, in this paper we use the 100-language WALS sample for a series of statistical analyses aimed at the detection of possible complexity trade-offs. To do that, we define four binary variables whose values can alternatively be "simple" or "complex", and those variables are built using information from certain features.[3]

The definitions of the abovementioned complexity variables are the following:

a) Phonology: A language is considered to be complex if it has more than 25 consonant

[2] This sample was used by us in previous work (Coloma, 2015). The complete list of languages is reproduced in appendix 1.

[3] The WALS features used are: 1A (Consonant inventories), 2A (Vowel quality inventories), 13A (Tone), 20A (Fusion of selected inflectional formatives), 26A (Prefixing vs. suffixing in inflectional morphology), 37A (Definite articles), 81A (Order of subject, object and verb), 119A (Nominal and locational predication) and 122A (Relativization on subjects).

phonemes, more than 6 vowel qualities, or uses tone as a distinctive phonological feature. This generates a division in which 60 languages are complex, and the remaining 40 languages are simple.

b) Morphology: A language is considered to be complex if it is polysynthetic, and simple if it is not. This implies that 32 languages in the sample are complex, and the remaining 68 ones are simple.

c) Syntax: A language is considered to be complex if it has no dominant word order for subject, object and verb, or if it uses relative pronouns to build relative clauses. Under this definition, 22 languages are complex and the remaining 78 ones are simple.

d) Lexicon: A language is considered to be complex if it has definite articles and uses different verbs for nominal and locational predication. This implies that 33 languages are complex, and the remaining 67 languages are simple.[4]

Table 1: Standard correlation coefficients between complexity variables

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.1400	1.0000		
Syntax	-0.0591	0.0497	1.0000	
Lexicon	-0.1650	-0.0711	-0.0647	1.0000

The easiest way to detect possible trade-offs between these binary complexity variables is to calculate standard (product-moment) correlation coefficients, like the ones that appear on table 1. In that table there are five negative correlation coefficients and one positive correlation coefficient, but none of them is statistically significant at a 5% probability level.[5]

[4] The value of each complexity variable for each language is reported in appendix 2, where "simple" is denoted as "0" and "complex" is denoted as "1".

[5] For any two variables whose correlation is calculated using 100 observations, correlation coefficients are statistically significant at a 5% probability level (i.e., the probability that the true correlation is zero is less than 5%) if they are greater than 0.2 in absolute value.

3. Simultaneous equation regressions

The standard correlation coefficients reported on table 1 are in all cases calculated using information that covers two variables for each coefficient. In this case, however, it is possible to consider that our measures of phonological, morphological, syntactic and lexical complexity are somehow interrelated, in the sense that the relationship between any pair of those measures can be influenced by the other complexity variables.

One way to model a situation like the one described in the previous paragraph is to build a system of simultaneous equations like the following:

$$Phonology = c(1) + c(2)*Morphology + c(3)*Syntax + c(4)*Lexicon \quad (1)\ ;$$
$$Morphology = c(5) + c(6)*Phonology + c(7)*\ Syntax + c(8)*\ Lexicon \quad (2)\ ;$$
$$Syntax = c(9) + c(10)*Phonology + c(11)*Morphology + c(12)*Lexicon \quad (3)\ ;$$
$$Lexicon = c(13) + c(14)*Phonology + c(15)*Morphology + c(16)*Syntax \quad (4)\ ;$$

where *Phonology*, *Morphology*, *Syntax* and *Lexicon* are the complexity variables defined for the 100-language WALS sample, whose values can either be equal to 0 (if the language is simple in the corresponding domain) or equal to 1 (if the language is complex in that domain). Additionally, coefficients *c(1)* to *c(16)* are the values of the parameters that relate each complexity measure with the other measures.

One easy way to estimate coefficients *c(1)* to *c(16)* is to run a set of four separate ordinary least-square (OLS) regressions. If we do that, we get the following results:

$$Phonology = 0.7286 - 0.1572*Morphology - 0.0749*Syntax - 0.1872*Lexicon \quad (5)\ ;$$
$$Morphology = 0.4300 - 0.1462*Phonology + 0.0389*Syntax - 0.0935*Lexicon \quad (6)\ ;$$
$$Syntax = 0.2649 - 0.0560*Phonology + 0.0313*Morphology - 0.0644*Lexicon \quad (7)\ ;$$
$$Lexicon = 0.4826 - 0.1749*Phonology - 0.0939*Morphology - 0.0804*Syntax \quad (8)\ .$$

With these results, it is possible to calculate new (partial) correlation coefficients, defined as the square roots of the products of the corresponding pairwise regression coefficients.[6] For example, for the relationship between phonological and morphological

[6] For a more thorough explanation of the concept of partial correlation, and the available alternatives for its calculation, see Prokhorov (2002).

complexity, this is equal to the square root of "–0.1572" (which is the regression coefficient of *Morphology* as a determinant of *Phonology*) times "–0.1462" (which is the regression coefficient of *Phonology* as a determinant of *Morphology*). As both regression coefficients are negative, we must assign a negative sign to the corresponding correlation coefficient (i.e., to the corresponding square root), whose value is "r = –0.1516". If we make similar calculations for all the possible pairwise relationships that appear in our system, we will have a set of numbers like the ones reported on table 2.

Table 2: Partial correlation coefficients between complexity variables

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.1516	1.0000		
Syntax	-0.0648	0.0349	1.0000	
Lexicon	-0.1809	-0.0937	-0.0720	1.0000

The procedure used to calculate the regression coefficients that appear in equations 5 to 8 (which is the basis for the calculation of the partial correlation coefficients reported on table 2) regresses each equation independently. However, if we use a truly simultaneous procedure in which the four equations are regressed at the same time, we can also use the correlation coefficients between the residuals of the different equations, and derive a new set of regression coefficients like the following:

$$Phonology = 0.8580 - 0.3149*Morphology - 0.1532*Syntax - 0.3743*Lexicon \quad (9)\ ;$$

$$Morphology = 0.5499 - 0.2928*Phonology + 0.0586*Syntax - 0.2032*Lexicon \quad (10)\ ;$$

$$Syntax = 0.3174 - 0.1146*Phonology + 0.0472*Morphology - 0.1324*Lexicon \quad (11)\ ;$$

$$Lexicon = 0.6414 - 0.3496*Phonology - 0.2040*Morphology - 0.1652*Syntax \quad (12)\ .$$

This new set of regression coefficients comes from a statistical method known as "seemingly unrelated regression" technique (SUR), originally proposed by Zellner (1962). This method is not very common in linguistics, but it is relatively widespread in other social sciences such as economics (where it is standard for applications like demand estimation). In this case, however, its use generates an important increase in the magnitude of the estimated negative partial correlation coefficients, which can now be

approximated by the numbers reported on table 3.

Table 3: Partial correlation coefficients using SUR

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.3036	1.0000		
Syntax	-0.1325	0.0526	1.0000	
Lexicon	-0.3618	-0.2036	-0.1479	1.0000

An additional variation that can be introduced is the elimination of the only positive correlation coefficient that we have obtained (which relates morphological and syntactic complexity), provided that its sign is counterintuitive and its absolute value (r = 0.0526) is small and statistically insignificant. If we do that, we can regress a new restricted system of equations, whose results (using SUR) are the following:

$$Phonology = 0.8591 - 0.3161*Morphology - 0.1563*Syntax - 0.3745*Lexicon \quad (13)\ ;$$

$$Morphology = 0.5686 - 0.2989*Phonology - 0.2099*Lexicon \quad (14)\ ;$$

$$Syntax = 0.3421 - 0.1263*Phonology - 0.1404*Lexicon \quad (15)\ ;$$

$$Lexicon = 0.6424 - 0.3498*Phonology - 0.2053*Morphology - 0.1671*Syntax \quad (16)\ .$$

The new partial correlation coefficients implied by this system of regression equations appear on table 4, in which three out of the five estimated coefficients (phonology vs. morphology, phonology vs. lexicon, and morphology vs. lexicon) are now statistically significant at a 5% probability level.

Table 4: Partial correlation coefficients using a restricted version of SUR

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.3074	1.0000		
Syntax	-0.1405	0.0000	1.0000	
Lexicon	-0.3620	-0.2076	-0.1532	1.0000

4. Instrumental variables

The logic behind the equations used to estimate the partial correlation coefficients between phonological, morphological, syntactic and lexical complexity has to do with the idea that those complexity levels come from a system that generates them as the outcome of some unified procedure. That procedure may consist of the interaction between several constraints such as the ones proposed by the synergetic linguistics' literature (e.g., Köhler 2005), or some kind of iterative learning mechanism like the one proposed by Smith, Kirby and Brighton (2003).

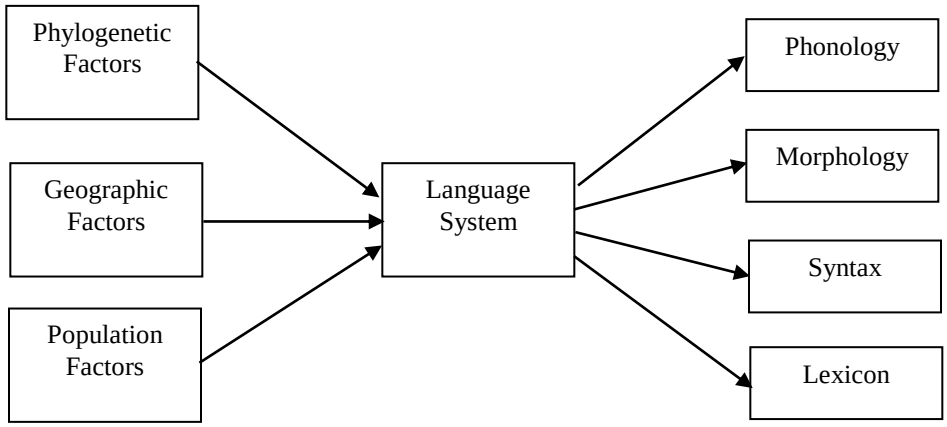
Those theoretical approaches share the common assumption that languages emerge in environments that can be influenced by a series of non-linguistic factors. Among those factors, the ones that are easier to analyze in empirical work are the geographic, phylogenetic and population characteristics of the different languages. For example, as any language originated in a certain point in space, it can be classified as belonging to a certain region or area (e.g., one of the six large macro-areas that the WALS defines).

The other major classification used by the linguistic literature is the phylogenetic one, which groups languages into families that share a common ancestor (e.g., Indo-European, Afro-Asiatic, Niger-Congo, etc.). A third element that we can use to classify languages is their relative size in terms of population, which is related to the geographic expansion that each language has had in history, and its alternative use as a first or second language by different people. Those characteristics have been used to analyze the relationship between language complexity and population, in papers such as Dahl (2011) or Bentz et al. (2015).

One general way to think about the relationship between linguistic and non-linguistic variables is to assume that the latter are part of the environment in which the former arise. This implies that linguistic variables may be influenced by non-linguistic factors, buy not the other way round. If we use this type of reasoning, we may represent our relationships by a graph like the one that appears in figure 1. In it we see that each linguistic variable (phonological, morphological, syntactic and lexical complexity) is the

outcome of a language system which has in turn been influenced by phylogenetic, geographic and population factors.

Figure 1: Relationships in a language system



The application of this view to the statistical explanation of the levels of language complexity implies the possibility of running a system of equations where those levels of complexity are the dependent variables, and the non-linguistic factors are the independent variables (on which the language complexity levels depend). In order to do that, we first need to encode the phylogenetic, geographic and population factors into numerical values, and the simplest way to do it is to create binary variables that take a value equal to one when a language belongs to a certain (geographic, phylogenetic or population) group, and zero otherwise.

Using the six WALS macro-areas and three additional divisions for those areas, we have created nine binary geographic variables that correspond to Eurasia, South East Asia, Africa, Papunesia, Australia, North America, Mesoamerica, the Amazon basin, and (the rest of) South America. Due to the fact that in the 100-language WALS sample there are relatively many observations that belong to three particular families (Austronesian, Indo-European, and Niger-Congo), we have also created three variables related to those phylogenetic factors. Finally, we have classified languages according to their relative

size, considering the ones with more than 5 million native speakers as "major", and the other ones as "minor".[7]

The next step in the estimation of the effect of non-linguistic factors on language variables was to run a system of OLS equations in which each of the four complexity measures used in the previous sections was regressed against the thirteen non-linguistic binary variables. As the sum of the geographic variables completely covers the whole sample of languages, we have used one region (Eurasia) as the default one (constant), and included the remaining geographic variables as explanatory variables. The results for each of the four regressions appear on table 5.

Table 5: Regression coefficients for the language complexity variables

Explanatory variables	Phonology	Morphology	Syntax	Lexicon
Constant	0.5919	0.2839	0.2471	0.3461
Africa	0.2445	-0.1906	-0.2413	0.3149
South East Asia	0.2117	-0.1674	-0.1817	0.2827
Papunesia	-0.1318	0.0834	-0.2281	0.0316
Australia	-0.5919	0.2876	0.0386	-0.2032
North America	0.0747	0.5495	0.2529	0.2373
Mesoamerica	0.0747	0.2161	-0.2471	-0.1794
South America	-0.4862	0.1720	-0.0157	0.3958
Amazon	0.0331	0.4661	0.0029	-0.3461
Austronesian	-0.4706	-0.2806	0.2621	0.4699
Indo-European	-0.2026	-0.1441	0.5814	0.1335
Niger-Congo	0.0626	-0.0334	0.0279	-0.2233
Major Language	0.2357	-0.1398	-0.0785	-0.3546

The outcome of this regression analysis shows results that are in line with received linguistic knowledge. We can see, for example, that Australian and Austronesian languages tend to have simpler phonologies, that North American languages tend to have more complex morphologies, that Indo-European languages tend to have more complex

[7] The 33 major languages in the 100-language WALS sample are the following: Mandarin, English, Spanish, Hindi, Arabic, Russian, Japanese, German, French, Indonesian, Korean, Turkish, Vietnamese, Persian, Kannada, Hausa, Burmese, Tagalog, Yoruba, Swahili, Oromo, Thai, Malagasy, Greek, Zulu, Quechua, Berber, Hebrew, Khalkha, Finnish, Guarani, Georgian and Hmong Njua. The remaining 67 ones are considered to be "minor languages". To see which languages belong to the different geographic and phylogenetic groups, see appendix 1.

levels of syntax, and that major languages tend to have a simpler lexicon (but a more complex phonology). These results may also be combined with the analysis that we performed in section 3, since the newly obtained regression coefficients can be the basis to build variables that are "instrumental" for that analysis.

The way to build these instrumental variables is to recover the predictions of the different regressions for each of the dependent variables of those regressions. With that we obtain four new variables (*Phonôlogy*, *Morphôlogy*, *Syntâx* and *Lexîcon*), which are linear combinations of the values of our thirteen binary non-linguistic variables (multiplied by their respective regression coefficients). These instrumental variables have the property that they can replace the original variables of the regression systems run in section 3, and are at the same time completely exogenous to those systems.

Instrumental variables are a useful device to solve a statistical problem known as the "endogeneity problem". This arises when we run a regression in which we know that both the dependent variable and (at least one of) the independent variables are somehow determined by the same mechanism. When this is the case, the obtained regression coefficients can be biased or inconsistent. If, however, we replace the endogenous independent variables by other variables that serve as exogenous instruments to approximate the value of those variables, then the estimation may turn less precise but more consistent and unbiased.[8]

In the system of equations introduced in section 3, all variables seem to be endogenous in the sense described in the previous paragraph. This is because they are at the same time dependent variables in one equation and independent variables in other equations, and all the relationships are supposed to be generated by the same mechanism. If we add the idea that such mechanism is somehow influenced by non-linguistic factors like the ones represented by the set of phylogenetic, geographic and population variables included in the regressions performed in this section, we can think of those variables as good candidates to act as exogenous instruments to replace the original (endogenous) linguistic variables.

[8] For a more complete explanation of the endogeneity problem, see Kennedy (2008), chapter 9.

The statistical method that uses a set of instruments to estimate instrumental variables, and then uses those instrumental variables to replace the original endogenous variables in a context of a simultaneous-equation regression estimation is known as "three-stage least squares" (3SLS). It was originally proposed by Zellner and Theil (1962), and is widely used in other social sciences such as economics (where it is standard for problems such as supply and demand estimation). If we use this method to run the system formed by equations 13 to 16, what we find is the following:

*Phonology = 1.1276 –0.5125*Morphôlogy –0.6073*Syntâx –0.6969*Lexîcon* (17) ;

*Morphology = 0.7413 –0.5560*Phonôlogy –0.2657* Lexîcon* (18) ;

Syntax = 0.5538 –0.4416 Phonôlogy –0.2088* Lexîcon* (19) ;

Lexicon = 0.7729 –0.5569 Phonôlogy –0.1947*Morphôlogy –0.2112*Syntâx* (20) .

With these regression coefficients, we can now derive new partial correlation coefficients, which are the ones that appear on table 6. There we can see that the five estimated coefficients are now negative and statistically significant at a 5% probability level, since all of them are higher than 0.2 in absolute value.

Table 6: Partial correlation coefficients using 3SLS

Variables	Phonology	Morphology	Syntax	Lexicon
Phonology	1.0000			
Morphology	-0.5338	1.0000		
Syntax	-0.5179	0.0000	1.0000	
Lexicon	-0.6230	-0.2275	-0.2100	1.0000

5. Concluding remarks

The analysis performed in this paper about possible complexity trade-offs in the 100-language WALS sample can be seen as a particular statistical exercise whose outcome is likely to change if we use other language samples or other definitions for the different types of language complexity. The message that we get from that analysis, however, is probably of a more general nature, since it seems to reconcile some contradictory results from previous literature.

In the very beginning, our analysis generates the standard result that, if we measure trade-offs using product-moment correlation coefficients between typological complexity measures, what we get is a set of statistically insignificant values which imply that language complexity trade-offs are either inexistent or unimportant (and this is equivalent to the conclusions of papers such as Shosted 2006). When we use non-linguistic variables related to phylogenetic, geographic and population factors, conversely, we obtain results that indicate that some complexity variables may indeed be influenced by those factors, and this seems to be in line with some contributions from sociolinguistic typology (e.g., Trudgill 2009).

What we do not get, if we restrict ourselves to standard correlation and regression techniques, is anything related to the logic behind the idea of language as a complex adaptive system, since that idea implies that language should evolve to be at the same time "compressed" (i.e., relatively simple and easy to learn) and "expressive" (i.e., relatively complex and capable to convey meanings for multiple concepts).[9] To reconcile these two requirements, we need to find some kind of trade-off between different levels of language complexity, such as the ones that typically appear in the literature that uses empirical measures of complexity (e.g., Fenk-Oczlon and Fenk 2008).

In one contribution that belongs to that literature (Coloma 2016) we got a result that shows that language complexity trade-offs seem to be more significant if we measure them using partial correlation coefficients instead of standard correlation coefficients, and they get even more significant if we use simultaneous-equation regression methods such as SUR. We therefore decided to apply the same logic to study the possible trade-offs between typological complexity measures, since simultaneous-equation regression methods have been designed to deal with statistical problems in which the different equations that we want to regress are generated by the same mechanism. And this is precisely the case here, where we are supposing that language complexity variables come from some kind of unified generating process.

But, as we also have variables related to non-linguistic factors that may influence

[9] For an interesting analysis of this dichotomy, see Kirby et al. (2015).

the language system from outside, we can use those variables to solve a statistical problem that simultaneous-equation models usually have, which is the endogeneity problem. To solve this we use non-linguistic factors to create instrumental variables, and then we use those instrumental variables as part of a 3SLS procedure. In this case, this can be seen as the statistical representation of a model in which non-linguistic factors are able to influence the system in which language is produced, and such system is in turn the one that generates the (interrelated) levels of complexity that correspond to its different sub-systems (i.e., phonology, morphology, syntax and lexicon).

When we did this, our results changed dramatically. Except for the coefficient that relates morphology and syntax, which is always insignificant, all the other correlation coefficients are negative and statistically significant when we estimate them using 3SLS. Moreover, their statistical significance increases when we move from standard to partial (OLS) coefficients, and the same occurs when we move from OLS to SUR, and from SUR to 3SLS coefficients.

This behavior may obey to different causes, but one plausible one is the idea that the statistical sophistications included in our calculations are related to an increasing consideration of the interactions between language complexity variables. When we only use standard correlation coefficients, those interactions are computed pairwise, while the calculation of partial correlation coefficients through an OLS procedure implies considering multiple interactions as well. Using the SUR method is in turn equivalent to introduce relationships between the errors that arise when we estimate the different complexity equations, whereas 3SLS implies considering the effect of non-linguistic factors (and their influence on the system that is producing the different levels of language complexity).

As a final conclusion, therefore, we can say that language complexity trade-offs may be more pervasive than what it seems when we measure them using simple statistical tools such as standard correlation coefficients or univariate regression equations. This is because there may be some interferences from other factors, whose effects have to be taken into account using more sophisticated statistical procedures. But that is indeed the

message implied by the theoretical approach that sees language as a complex adaptive system, and the use of simultaneous-equation regression models can be a way to interpret the available data which is compatible with that approach.

References

Beckner, Clay, Richard Blythe, Joan Bybee, Morten Christiansen, William Croft, Nick Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman & Tom Schoenemann (2009). "Language Is a Complex Adaptive System". *Language Learning*, vol 59, suppl 1, pp 1-26.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill & Paula Buttery (2015). "Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms". *PLOS One*, vol 10(6), e0128254.

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

Coloma, Germán (2015). "Efectos de compensación entre indicadores de la complejidad de los idiomas"; Working Paper No. 569. Buenos Aires, CEMA University.

Coloma, Germán (2016). "The Existence of Negative Correlation Between Linguistic Measures Across Languages". *Corpus Linguistics and Linguistic Theory*, forthcoming.

Dahl, Osten (2011). "Are Small Languages More or Less Complex than Big Ones?" *Linguistic Typology*, vol 15, pp 171-175.

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*, pp 43-65. Amsterdam: John Benjamins.

Kennedy, Peter (2008). *A Guide to Econometrics*, 6th edition. New York: Wiley.

Kirby, Simon, Monica Tamariz, Hannah Cornish y Kenny Smith (2015). "Compression and Communication in the Cultural Evolution of Linguistic Structure". *Cognition*, vol 141, pp 87-102.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik: Stuktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, Reinhard (1987). "System Theoretical Linguistics". *Theoretical Linguistics*, vol 14, pp 241-257.

Köhler, Reinhard (2005). “Synergetic Linguistics”. In G. Altmann, R. Köhler & R. Piotrowski (eds.), *Quantitative Linguistics: An International Handbook*, pp 760-774. Berlin: De Gruyter.

Nettle, Daniel (1995). “Segmental Inventory Size, Word Length and Communicative Efficiency”. *Linguistics*, vol 33, pp 359-367.

Nichols, Johanna (2009). “Linguistic Complexity: A Comprehensive Definition and Survey”. In G. Sampson, D. Gil & P. Trudgill (eds.), *Language Complexity as an Evolving Variable*, pp 110-125. Oxford: Oxford University Press.

Prokhorov, A. V. (2002). “Partial Correlation Coefficient”. In M. Hazewinkel (ed.), *Encyclopedia of Mathematics*. New York: Springer.

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

Smith, Kenny, Simon Kirby & Henry Brighton (2003). “Iterated Learning: A Framework for the Emergence of Language”. *Artificial Life*, vol 9, pp 371-386.

Trudgill, Peter (2009). “Sociolinguistic Typology and Complexification”. In G. Sampson, D. Gil & P. Trudgill (eds.), *Language Complexity as an Evolving Variable*, pp 98-109. Oxford: Oxford University Press.

Zellner, Arnold (1962). “An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias”. *Journal of the American Statistical Association*, vol 57, pp 348–368.

Zellner, Arnold & Henri Theil (1962). “Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations”. *Econometrica*, vol 30, pp 54–78.

Appendix 1: List of languages in the WALS sample

Code	Language	Region	Family
1	Abkhaz	Eurasia	Northwest Caucasian
2	Acoma	North America	Keresan
3	Alamblak	Papunesia	Sepik
4	Amele	Papunesia	Trans-New Guinea
5	Apurina	Amazonia	Arawakan
6	Arabic (Egyptian)	Eurasia	Afro-Asiatic
7	Arapesh (Mountain)	Papunesia	Kombio
8	Asmat	Papunesia	Trans-New Guinea
9	Bagirmi	Africa	Nilo-Saharan
10	Barasano	Amazonia	Tucanoan
11	Basque	Eurasia	Vasconic
12	Berber (Middle Atlas)	Africa	Afro-Asiatic
13	Burmese	South East	Sino-Tibetan
14	Burushaski	Eurasia	Burushaskian
15	Canela-Kraho	Amazonia	Macro-Ge
16	Chamorro	Papunesia	Austronesian
17	Chukchi	Eurasia	Chukotkan
18	Cree (Plains)	North America	Algic
19	Daga	Papunesia	Dagan
20	Dani (Lower Valley)	Papunesia	Trans-New Guinea
21	English	Eurasia	Indo-European
22	Fijian	Papunesia	Austronesian
23	Finnish	Eurasia	Uralic
24	French	Eurasia	Indo-European
25	Georgian	Eurasia	Kartvelian
26	German	Eurasia	Indo-European
27	Gooniyandi	Australia	Bunuban
28	Grebo	Africa	Niger-Congo
29	Greek (Modern)	Eurasia	Indo-European
30	Greenlandic (West)	Eurasia	Eskimo-Aleut
31	Guarani	South America	Tupian
32	Hausa	Africa	Afro-Asiatic
33	Hebrew (Modern)	Eurasia	Afro-Asiatic
34	Hindi	Eurasia	Indo-European
35	Hixkaryana	Amazonia	Cariban
36	Hmong Njua	South East	Hmong-Mien
37	Imonda	Papunesia	Border
38	Indonesian	Papunesia	Austronesian
39	Jakaltek	Mesoamerica	Mayan

40	Japanese	Eurasia	Japonic
41	Kannada	Eurasia	Dravidian
42	Karok	North America	Karokian
43	Kayardild	Australia	Tangkic
44	Kewa	Papunesia	Trans-New Guinea
45	Khalkha	Eurasia	Altaic
46	Khoekhoe	Africa	Khoisan
47	Kiowa	North America	Tanoan
48	Koasati	North America	Muskogean
49	Korean	Eurasia	Koreanic
50	Koyraboro Senni	Africa	Nilo-Saharan
51	Krongo	Africa	Kaduglian
52	Kutenai	North America	Salish
53	Lakhota	North America	Siouan
54	Lango	Africa	Nilo-Saharan
55	Lavukaleve	Papunesia	East Papuan
56	Lezgian	Eurasia	Nakh-Daghestanian
57	Luvale	Africa	Niger-Congo
58	Makah	North America	Wakashan
59	Malagasy	Africa	Austronesian
60	Mandarin	South East	Sino-Tibetan
61	Mangarrayi	Australia	Mangarrayian
62	Mapudungun	South America	Araucanian
63	Maricopa	North America	Hokan
64	Martuthunira	Australia	Pama-Nyungan
65	Maung	Australia	Iwaidjan
66	Maybrat	Papunesia	West Papuan
67	Meithei	South East	Sino-Tibetan
68	Mixtec (Chalcatongo)	Mesoamerica	Oto-Manguean
69	Ngiyambaa	Australia	Pama-Nyungan
70	Oneida	North America	Iroquoian
71	Oromo (Harar)	Africa	Afro-Asiatic
72	Otomi (Mezquital)	Mesoamerica	Oto-Manguean
73	Paiwan	Papunesia	Austronesian
74	Persian	Eurasia	Indo-European
75	Piraha	Amazonia	Mura
76	Quechua (Imbabura)	South America	Quechuan
77	Rama	Mesoamerica	Chibchan
78	Rapanui	Papunesia	Austronesian
79	Russian	Eurasia	Indo-European
80	Sango	Africa	Niger-Congo
81	Sanuma	Amazonia	Yanomam

82	Slave	North America	Na-Dene
83	Spanish	Eurasia	Indo-European
84	Supyire	Africa	Niger-Congo
85	Swahili	Africa	Niger-Congo
86	Tagalog	Papunesia	Austronesian
87	Thai	South East	Tai-Kadai
88	Tiwi	Australia	Tiwian
89	Tukang Besi	Papunesia	Austronesian
90	Turkish	Eurasia	Altaic
91	Vietnamese	South East	Austro-Asiatic
92	Warao	South America	Waraoan
93	Wari	Amazonia	Chapacuran
94	Wichita	North America	Caddoan
95	Wichi	South America	Matacoan
96	Yagua	Amazonia	Peba-Yaguan
97	Yaqui	Mesoamerica	Uto-Aztecan
98	Yoruba	Africa	Niger-Congo
99	Zoque (Copainala)	Mesoamerica	Mixe-Zoque
100	Zulu	Africa	Niger-Congo

Appendix 2: Complexity variables

Code	Language	Phonology	Morphology	Syntax	Lexicon
1	Abkhaz	1	1	0	1
2	Acoma	1	1	1	0
3	Alamblak	1	0	0	1
4	Amele	0	1	0	0
5	Apurina	0	1	0	0
6	Arabic (Egyptian)	1	0	0	0
7	Arapesh (Mountain)	1	0	0	1
8	Asmat	0	1	0	0
9	Bagirmi	1	0	0	1
10	Barasano	1	0	1	0
11	Basque	0	0	0	1
12	Berber	1	0	0	0
13	Burmese	1	0	0	0
14	Burushaski	1	0	0	0
15	Canela-Kraho	1	1	0	0
16	Chamorro	0	0	0	1
17	Chukchi	0	1	1	0
18	Cree (Plains)	0	0	1	1
19	Daga	0	1	0	0

20	Dani (Lower Valley)	1	1	0	0
21	English	1	0	1	0
22	Fijian	0	0	1	0
23	Finnish	1	0	1	0
24	French	1	0	1	0
25	Georgian	1	0	1	0
26	German	1	0	1	0
27	Gooniyandi	0	0	1	0
28	Grebo	1	0	0	0
29	Greek (Modern)	0	0	1	0
30	Greenlandic (West)	0	1	0	0
31	Guarani	0	1	0	1
32	Hausa	1	0	0	1
33	Hebrew (Modern)	0	0	0	0
34	Hindi	1	0	0	0
35	Hixkaryana	0	1	0	0
36	Hmong Njua	1	0	0	0
37	Imonda	1	0	0	0
38	Indonesian	0	0	0	1
39	Jakaltek	1	1	0	0
40	Japanese	1	0	0	0
41	Kannada	1	0	0	0
42	Karok	1	1	1	1
43	Kayardild	0	1	1	0
44	Kewa	1	0	0	0
45	Khalkha	1	0	0	0
46	Khoekhoe	1	0	0	1
47	Kiowa	1	1	0	0
48	Koasati	1	1	0	1
49	Korean	1	0	0	0
50	Koyraboro Senni	0	0	0	1
51	Krongo	1	0	0	0
52	Kutenai	1	1	1	1
53	Lakhota	0	0	0	1
54	Lango	1	0	0	0
55	Lavukaleve	0	0	0	1
56	Lezgian	1	0	0	0
57	Luvale	1	0	0	0
58	Makah	1	1	0	1
59	Malagasy	1	0	0	1
60	Mandarin	1	0	0	0
61	Mangarrayi	0	1	0	0

62	Mapudungun	0	1	0	1
63	Maricopa	0	1	0	0
64	Martuthunira	0	0	0	1
65	Maung	0	1	0	0
66	Maybrat	0	0	0	1
67	Meithei	1	0	0	1
68	Mixtec	1	0	0	0
69	Ngiyambaa	0	0	0	0
70	Oneida	1	1	1	1
71	Oromo (Harar)	1	0	0	0
72	Otomi (Mezquital)	1	0	0	0
73	Paiwan	0	0	1	1
74	Persian	0	0	0	0
75	Piraha	1	1	0	0
76	Quechua (Imbabura)	1	0	0	0
77	Rama	0	0	0	0
78	Rapanui	0	0	0	1
79	Russian	1	0	1	0
80	Sango	1	0	0	1
81	Sanuma	1	1	0	0
82	Slave	1	1	0	0
83	Spanish	0	0	1	1
84	Supyire	1	0	0	1
85	Swahili	1	0	0	0
86	Tagalog	0	0	0	0
87	Thai	1	0	0	0
88	Tiwi	0	1	0	0
89	Tukang Besi	0	0	0	1
90	Turkish	1	0	0	0
91	Vietnamese	1	0	0	1
92	Warao	0	0	1	0
93	Wari	0	0	0	0
94	Wichita	0	1	1	0
95	Wichi	0	0	0	1
96	Yagua	1	1	1	0
97	Yaqui	1	1	0	0
98	Yoruba	1	0	0	0
99	Zoque (Copainala)	0	1	0	1
100	Zulu	1	0	0	0
Tot	**Total**	60	32	22	33