GLOTTOMETRICS 35, 2016, 49-63

AN OPTIMIZATION MODEL OF GLOBAL LANGUAGE COMPLEXITY

Germán Coloma
CEMA University, Buenos Aires, Argentina
gcoloma@cema.edu.ar

Abstract. In this paper we develop a theoretical model of global language complexity, based on a constrained optimization approach. We assume that language is a system that chooses different levels of complexity for its different domains (i.e., phonology, morphology, syntax, vocabulary) in order to minimize a global complexity function subject to an expressivity constraint (which also depends on non-linguistic variables related to geographic, phylogenetic and demographic factors). The model is illustrated with the aid of a dataset based on a short text translated into 50 languages, for which global complexity is measured using a version of Kolmogorov complexity. That dataset is used to run simultaneous-equation regressions, which represent different relationships between language complexity measures.

***Keywords**: language complexity, optimization, Kolmogorov complexity, simultaneous-equation regression.*

1. Introduction

The literature about global language complexity is relatively vast and diverse. On one hand, there is a considerable amount of theoretical literature that has dealt with topics such as the definition of language complexity (e.g., Kusters 2003, Miestamo 2008, Culicover 2013) and its determinants (e.g., McWhorther 2001, Balasubrahmanyan & Naranan 2002, Hawkins 2004, Trudgill 2009). On the other hand, there is a good deal of empirical work that has either analyzed the relationship between complexity measures (e.g., Nettle 1995, Fenk-Oczlon & Fenk 2005, Shosted 2006, Sinnemäki 2008) or the relationship between those measures and other (non-linguistic) variables (e.g., Hay & Bauer 2007, Atkinson 2011).

The theoretical literature has also developed models assuming that language is a system, and that its behavior is guided by a hidden process which tries to achieve some desired objective. Among the main contributions to that literature we can mention Beckner et al. (2009), which states that language is a complex adaptive system whose structures emerge from interrelated patterns of experience, social interactions and cognitive mechanisms. Another group of studies in a similar line are the ones related to the concept of "synergetic linguistics" (e.g., Köhler 2005), for which language is a self-organizing and self-regulating system whose properties come from the interaction of several constitutive, forming and control requirements.

Part of that theoretical literature has explored the possibility of explaining the behavior of the language system through an optimization model (e.g., Ke, Ogura & Wang 2003, Ferrer-i-Cancho 2014, Futrell, Mahowald & Gibson 2015). However, we have not found any example

from that literature in which the model used is directly related to complexity minimization, and this is probably the main contribution of the current article. The model that we develop here is, nevertheless, well known in other social sciences such as economics (e.g., Chiang & Wainwright 2005), where cost minimization is a standard approach.

The main challenge for using a model like this is probably the fact that global language complexity is a rather indefinable concept, and it is therefore very hard to measure. The theoretical literature that has sought to find results related to its determinants (e.g., Hawkins 2004, Culicover 2013) has in general ended up with the conclusion that language complexity had better be studied using concepts that are applicable to specific situations (e.g., markedness, economy, efficiency). In the empirical literature, however, there is a measure derived from information theory that could represent the global complexity of a text. That measure is Kolmogorov complexity (Kolmogorov 1963), and it has been used by some authors in different linguistic settings (e.g., Juola 2008, Ehret & Szmrecsányi 2015).

Kolmogorov complexity can be defined as the length of the smallest algorithm required to generate a certain string of characters (Li & Vitányi 1997). Although in general it is formally incomputable, it can be approximated by the size of a compressed text file that comes from another (original) file. The ratio between the sizes of the two files, therefore, can be seen as an empirical measure of the global complexity of the text to which those files refer to, since the possibility of compressing the original file into a smaller one is directly linked to a series of characteristics (e.g., letter inventory, letter repetition, morpheme repetition, word repetition, clause length) that signal the complexity of the text.

In the following pages we will develop a model in which we assume that global language complexity is measurable (for example, by computing the Kolmogorov complexity of a representative text) and that it depends on several partial complexity variables (which can also be measured). We will also assume that those partial complexity variables are somehow "chosen" by the language under analysis in order to minimize global complexity, but that they are also influenced by non-linguistic variables related to phylogenetic, geographic and demographic factors. Those factors can also be important to determine language "expressivity", i.e., the capacity of a language to discriminate between possible alternative referents for a certain expression (Kirby et al. 2015). That expressivity will also depend on the different language domains involved in the production, transmission and decoding of linguistic messages (e.g., phonology, morphology, syntax and vocabulary).

Our model will be illustrated with an example based on data from a short text for which we have translations to 50 different languages. With those translations we compute several complexity measures (including Kolmogorov complexity) and build a dataset in which those measures are seen as the variables of the empirical version of our model. As the languages belong to different families and regions, and are spoken by different numbers of people, we can make use of that diversity to build three additional (categorical) variables. With all that we proceed to estimate the parameters implicit in our theoretical model, using a statistical procedure of simultaneous-equation regressions known as "three-stage least squares" (Zellner & Theil 1962).

2. Theoretical model

Let us assume that we can measure the global complexity of a language by a numerical continuous variable "g". Let us suppose, moreover, that the value of that variable is an increasing function "C" of several partial complexity variables related to different language domains (e.g., phonology, morphology, syntax, vocabulary). Let us now assume that those partial complexity

variables are themselves numerical and continuous, and can be associated to “g” in the following way:

$$g = C(p, m, s, v) \quad (1)$$

where “p”, “m”, “s” and “v” may represent, for example, the phonological, morphological, syntactic and lexical complexity of language.

In a context like this, global complexity can be seen as a measure of the effort that speakers have to exert in order to use the language under analysis. Therefore, the smaller the value of “g”, the less costly a language is to be used by its speakers. But as language has to express meanings associated to its different components (i.e., to its words, clauses, texts, etc.), then its partial complexity levels can also be positively associated to its expressivity (through a function “E”, which will be increasing in “p”, “m”, “s” and “v”).

Following the ideas that appear in the literature about language as a complex adaptive system, we can think of the process of language evolution and transmission as an attempt to choose optimal levels for “p”, “m”, “s” and “v”, which simultaneously minimize “C” and maximize “E”. But this trade-off between opposing objectives can be influenced by other variables, such as phylogenetic, geographic and demographic factors (“pg”, “gg”, “dg”). One possible way to introduce those factors is to suppose that they operate as a determinant of the level of expressivity that a language must possess, through a restriction “R” (which integrates them into a single function). If that is the case, we can think of an “expressivity constraint” that can be written in the following way:

$$R(pg, gg, dg) = E(p, m, s, v) \quad (2) .$$

If “R” is a constraint for the level of “E”, and its determinants are exogenous to the language system, then our problem of choosing the optimal levels of “p”, “m”, “s” and “v” is somehow simplified, since it can be converted into one where we minimize “g” subject to the constraint stated in (2). If “C” and “E” are both continuous and differentiable in “p”, “m”, “s” and “v”, that problem can be solved using a standard optimization technique known as the “Lagrange method”. This method implies writing a Lagrangean function “L”, which is defined as follows:

$$L = C(p, m, s, v) + \lambda \cdot [R(pg, gg, dg) - E(p, m, s, v)] \quad (3)$$

and then finding the values of “p”, “m”, “s” and “v” for which the corresponding partial derivatives of “L” are equal to zero. These equalities are the “first-order conditions” of the problem, and can be written as:

$$\frac{\partial L}{\partial p} = \frac{\partial C}{\partial p} - \lambda \cdot \frac{\partial E}{\partial p} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C / \partial p)}{(\partial E / \partial p)} \quad (4)$$

$$\frac{\partial L}{\partial m} = \frac{\partial C}{\partial m} - \lambda \cdot \frac{\partial E}{\partial m} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C / \partial m)}{(\partial E / \partial m)} \quad (5)$$

$$\frac{\partial L}{\partial s} = \frac{\partial C}{\partial s} - \lambda \cdot \frac{\partial E}{\partial s} = 0 \quad \rightarrow \quad \lambda = \frac{(\partial C / \partial s)}{(\partial E / \partial s)} \quad (6)$$

$$\frac{\partial L}{\partial v} = \frac{\partial C}{\partial v} - \lambda \cdot \frac{\partial E}{\partial v} = 0 \qquad \rightarrow \qquad \lambda = \frac{(\partial C / \partial v)}{(\partial E / \partial v)} \qquad (7)\ .$$

In both the Lagrangean function and in its first-order conditions there is an additional "artificial variable" (λ), which is known as the "Lagrange multiplier" of the problem's constraint. This variable plays the role of converting the units in which the constraint is expressed (which in our case would be "expressivity units") into the units in which the objective function is expressed (i.e., complexity units). Due to that conversion, the first-order conditions can be stated as equations that relate infinitesimal changes in complexity with infinitesimal changes in expressivity, and establish optimal ratios between those changes.

Another role that the Lagrange multiplier plays is to include the fulfillment of the constraint as an additional first-order condition of the problem. This is due to the fact that, in order to minimize "C" subject to "R = E", we also need that:

$$\frac{\partial L}{\partial \lambda} = R(pg, gg, po) - E(p, m, s, v) = 0 \qquad \rightarrow \qquad R(pg,gg,dg) = E(p,m,s,v) \qquad (8)$$

and this last equation is included, together with equations (4) to (7), in a system whose solution is the one that determines the optimal values of "p", "m", "s" and "v" (and "λ").[1]

One relatively straightforward way to solve this system of equations is to use (4), (5), (6) and (7) to find the optimal relationships between each pair of partial complexity variables. By doing that, it is possible to express any complexity variable as a function of any other complexity variable (e.g., "m = m(p)", "s = s(p)", "s = s(m)", etc.). Making use of that possibility, we can replace those functions into (8), and write something like the following:

$$R(pg,gg,po) = E(p,m(p),s(p),v(p)) \qquad \rightarrow \qquad p = p(pg, gg, dg) \qquad (9)$$

$$R(pg,gg,po) = E(p(m),m,s(m),v(m)) \qquad \rightarrow \qquad m = m(pg, gg, dg) \qquad (10)$$

$$R(pg,gg,po) = E(p(s),m(s),s,v(s)) \qquad \rightarrow \qquad s = s(pg, gg, dg) \qquad (11)$$

$$R(pg,gg,po) = E(p(v),m(v),s(v),v) \qquad \rightarrow \qquad v = v(pg, gg, dg) \qquad (12)\ .$$

What equations (9) to (12) give us is actually the solution to our optimization problem. Each partial complexity variable is expressed as a function of the different phylogenetic, geographic and demographic factors that influence the language system under analysis, and can be inserted into equation (1) in order to get the minimum level of global complexity which is compatible with the fulfillment of the constraint stated in equation (2). By doing that, we obtain the following:

$$g = C(p(pg,gg,dg),\ m(pg,gg,dg),\ s(pg,gg,dg),\ v(pg,gg,dg)) \qquad (13)$$

which is an expression in which "g" is equated to a function that will ultimately depend on the actual levels of the non-linguistic variables. The whole process implied by this optimization model can therefore be represented by a graph like the one that appears in Figure 1.

As we can see, the idea behind this model is that the non-linguistic factors (i.e., phylogenetic, geographic and demographic), which influence the environment where language systems

[1] For a more complete explanation of this procedure, see Sundaram (1996), chapter 5.

operate, have an effect on the way in which those systems choose the characteristics of their different domains (i.e., phonology, morphology, syntax and vocabulary). But as those characteristics are determined simultaneously, then their implied levels of partial complexity are related to each other, and they all have an impact on the global complexity of the system.

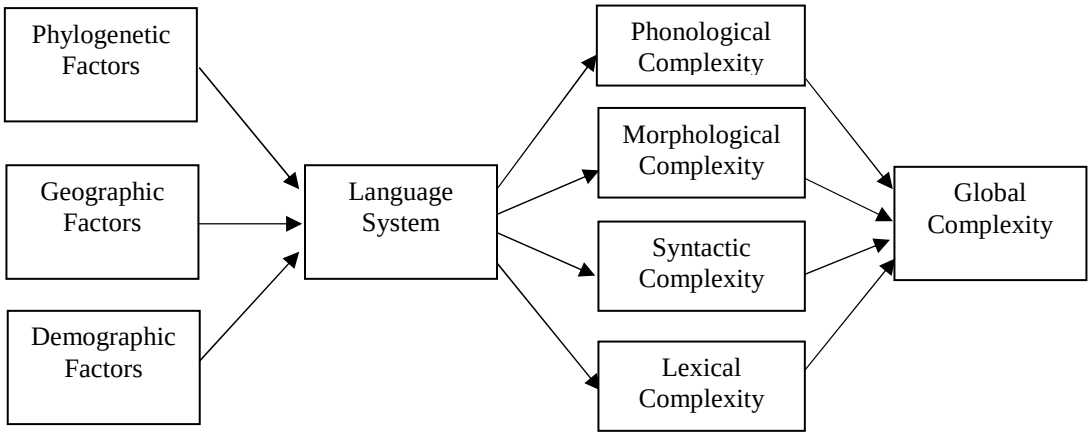

Figure 1. Language complexity model

3. Description of the data

In order to apply the theoretical model described in the previous section to an empirical example, we will use the same dataset that was previously employed in Coloma (2015, 2016), whose source is a series of articles published in IPA (1999) and in the *Journal of the International Phonetic Association*. It consists of a sample of 50 languages for which we have a version of the same text (the fable known as "The North Wind and the Sun"), on which we define different phonological, morphological, syntactic and lexical measures of complexity.[2] Those measures are the following:

Phonological inventory (*INV*): It is an index that consists of the sum of the number of consonant and vowel phonemes in each language, modified by the number of distinctive tones that such language possesses, and by the possible existence of distinctive levels of stress. This index is defined as:

$$INV = Consonants + Vowels*(Tones+Stress) \quad ;$$

where *Consonants*, *Vowels* and *Tones* are numerical variables, and *Stress* is a binary variable that takes a value equal to one when stress is distinctive in a certain language (and zero otherwise).

Phoneme/word ratio (*PWR*): It is defined as the ratio between the total number of phonemes of "The North Wind and the Sun" text in each language, and the corresponding total number of words in that text.

Word/clause ratio (*WCR*): It is defined as the ratio between the total number of words of "The North Wind and the Sun" text in each language, and the corresponding total number of clauses in that text.

[2] The list of languages and their complexity levels are reported in the appendix.

Type/token ratio (*TTR*): It is defined as the ratio between the number of different words (types) of "The North Wind and the Sun" text in each language, and the total number of words (tokens) in that text.

To measure the global complexity of the texts under analysis we will use Kolmogorov complexity (*KC*). This will be defined as the ratio between the size of a compressed file (which contains the "The North Wind and the Sun" in a certain language) and the size of the original version of that file, both of them measured in bytes. Compression was made using the program "7zip", version 4.32.

Another set of variables that we need for our empirical exercise is the one related to non-linguistic factors. This consists of one geographic variable, one phylogenetic variable and one demographic variable, all of which are categorical. The values of the geographic variable represent 10 different regions of the world, and each of them encompasses between 4 and 6 languages from our sample. The regions are North America, South America, Northern Europe, Southern Europe, Northern Africa, Southern Africa, West Asia, Central Asia, East Asia and the Pacific.

Additionally, the 50 languages in the sample belong to 24 different families, some of which are represented by more than one language. Those families are Indo-European (IE, 13 languages), Afro-Asiatic (AA, 5 languages), Niger-Congo (NC, 4 languages), Sino-Tibetan (ST, 3 languages), Altaic (Alt, 3 languages), Austronesian (Aus, 2 languages), Nilo-Saharan (NS, 2 languages), Oto-Manguean (OM, 2 languages), plus two languages that can generically be referred to as "Amazonian" (Amaz). The remaining 14 languages are grouped into another category named "Other families".[3]

The demographic variable, finally, divides languages into three categories: "large languages", "medium languages" and "small languages" (according to the number of speakers that the different languages possess). The group of large languages is constituted by the following 12 cases: Mandarin, English, Spanish, Hindi, Arabic, Portuguese, Russian, Japanese, Bengali, German, French and Malay. Correspondingly, the ones that belong to the category of "small languages" (less than 1 million speakers) are the following: Apache, Arrernte, Basque, Chickasaw, Dinka, Irish, Mapudungun, Nara, Sahaptin, Sandawe, Seri, Shiwilu, Tausug, Trique and Yine. The remaining 23 languages in the sample are considered to be "medium-sized".

The main descriptive statistics of this database are summarized in Table 1, in which we find the average values of *INV*, *PWR*, *WCR*, *TTR* and *KC* for each group of languages. In that table we can see, for example, that East Asian and Oto-Manguean languages tend to be phonologically more complex, and that Amazonian languages tend to have higher phoneme/word ratios. The higher word/clause ratios, conversely, appear in Indo-European languages (especially in the Northern European ones), while the highest type/token ratios seem to occur in West Asia and in Sino-Tibetan languages. Finally, Kolmogorov complexity is higher in Southern Europe and East Asia, and in Sino-Tibetan and Altaic languages.

Another set of descriptive statistics that could be useful to analyze our complexity measures is the one formed by the correlation coefficients between the different variables. Those coefficients are reported in Table 2, in which we see that there are several partial complexity measures that display relatively significant negative correlation coefficients between themselves. The most important one is the coefficient between *PWR* and *WCR* (r = -0.7265), followed by the one between *WCR* and *TTR* (r = -0.5046), and by the one between *INV* and *PWR* (r = -0.3720). Conversely, Kolmogorov complexity exhibits relatively large positive correlation coefficients

[3] To see which language belongs to each family, see the table included in the appendix.

with *TTR* and *INV* ("r = 0.3639" and "r = 0.2543"), and small negative correlation coefficients with *PWR* and *WCR* ("r = -0.1395" and "r = -0.1108").

Table 1
Descriptive statistics for complexity variables

Category / Variable	INV	PWR	WCR	TTR	KC
Northern Europe	44.20	3.964	12.53	0.6336	0.7203
Southern Europe	33.60	4.086	11.93	0.6105	0.7490
Northern Africa	42.75	4.787	11.04	0.6417	0.6978
Southern Africa	49.00	4.469	10.98	0.6296	0.7002
North America	49.83	5.115	9.22	0.5938	0.6851
South America	25.00	7.076	7.58	0.6541	0.6357
West Asia	33.50	5.854	9.18	0.7634	0.7417
Central Asia	36.00	5.060	11.33	0.7067	0.6573
East Asia	68.50	4.709	10.32	0.6865	0.7911
Pacific	26.25	5.530	8.80	0.5838	0.6450
Indo-European	38.38	4.259	12.28	0.6555	0.7134
Afro-Asiatic	39.20	5.277	10.40	0.7164	0.6956
Niger-Congo	43.00	4.298	11.15	0.6231	0.7231
Sino-Tibetan	66.00	5.128	8.23	0.7469	0.8291
Altaic	31.00	6.009	8.52	0.7132	0.8103
Austronesian	22.00	5.592	9.63	0.5489	0.6252
Nilo-Saharan	46.50	4.157	11.76	0.5469	0.6792
Oto-Manguean	68.00	3.557	10.18	0.6173	0.7749
Amazonian	23.00	8.457	6.91	0.6904	0.5874
Other families	45.50	5.361	9.44	0.6302	0.6787
Large languages	37.67	4.439	11.15	0.6507	0.7153
Medium languages	44.48	5.048	10.30	0.6873	0.7315
Small languages	42.60	5.438	9.65	0.5992	0.6661
Average	**42.28**	**5.019**	**10.31**	**0.6521**	**0.7080**

Table 2
Correlation coefficients between complexity variables

Complexity variable	INV	PWR	WCR	TTR	KC
Phonological inventory	1.0000				
Phoneme/word ratio	-0.3720	1.0000			
Word/clause ratio	0.2136	-0.7265	1.0000		
Type/token ratio	-0.0428	0.5581	-0.5046	1.0000	
Kolmogorov complexity	0.2543	-0.1395	-0.1108	0.3639	1.0000

4. Empirical estimation

The dataset that we described in section 3 can be used to perform an estimation of the model developed in section 2. In order to do that, we first need to define which empirical variables will be used to approximate the theoretical variables of the model, and which functional forms can approximate the relationships that that model displays.

One obvious possibility is to use *INV*, *PWR*, *WCR* and *TTR* as proxies for "p", "m", "s" and "v", respectively. An easy way to use them to write expressions for the functions "C" and "E" is to suppose that the first of those functions is linear, and that the second one is log-linear. This implies that both functions will depend on four variables and four parameters each, and they can be written as:

$$C = c1 \cdot INV + c2 \cdot PWR + c3 \cdot WCR + c4 \cdot TTR \quad (14) ;$$

$$E = a1 \cdot \ln(INV) + a2 \cdot \ln(PWR) + a3 \cdot \ln(WCR) + a4 \cdot \ln(TTR) \quad (15) ;$$

where *c1*, *c2*, *c3* and *c4* are the parameters of the complexity function, and *a1*, *a2*, *a3* and *a4* are the parameters of the expressivity function.

In order to perform a statistical estimation of "C", it is straightforward to assume that global complexity can be approximated by the value of *KC*. This implies that parameters *c1*, *c2*, *c3* and *c4* are going to be the results of a procedure in which *KC* is regressed as a linear function of *INV*, *PWR*, *WCR* and *TTR*. The estimation of "E", conversely, is considerably more cumbersome, since we do not have any empirical variable that can easily be associated to a measure of expressivity. What we can do, instead, is to work with the first-order conditions of the theoretical optimization problem described in section 2, and write them in the following way:

$$\frac{\partial C/\partial p}{\partial E/\partial p} = \frac{\partial C/\partial m}{\partial E/\partial m} = \frac{\partial C/\partial s}{\partial E/\partial s} = \frac{\partial C/\partial v}{\partial E/\partial v} \quad \rightarrow \quad \frac{c1}{a1/INV} = \frac{c2}{a2/PWR} = \frac{c3}{a3/WCR} = \frac{c4}{a4/TTR} \quad (16) .$$

As those relationships imply equality signs, it is possible to write equations that relate complexity variables in pairs. Those pairs are the following:

$$INV = \frac{a1}{a2} \cdot \frac{c2}{c1} \cdot PWR \ ; \qquad INV = \frac{a1}{a3} \cdot \frac{c3}{c1} \cdot WCR \ ; \qquad INV = \frac{a1}{a4} \cdot \frac{c4}{c1} \cdot TTR \quad (17)$$

$$PWR = \frac{a2}{a1} \cdot \frac{c1}{c2} \cdot INV \ ; \qquad PWR = \frac{a2}{a3} \cdot \frac{c3}{c2} \cdot WCR \ ; \qquad PWR = \frac{a2}{a4} \cdot \frac{c4}{c2} \cdot TTR \quad (18)$$

$$WCR = \frac{a3}{a1} \cdot \frac{c1}{c3} \cdot INV \ ; \qquad WCR = \frac{a3}{a2} \cdot \frac{c2}{c3} \cdot PWR \ ; \qquad WCR = \frac{a3}{a4} \cdot \frac{c4}{c3} \cdot TTR \quad (19)$$

$$TTR = \frac{a4}{a1} \cdot \frac{c1}{c4} \cdot INV \ ; \qquad TTR = \frac{a4}{a2} \cdot \frac{c2}{c4} \cdot PWR \ ; \qquad TTR = \frac{a4}{a3} \cdot \frac{c3}{c4} \cdot WCR \quad (20) .$$

The equations that appear in (17), (18), (19) and (20) can also be added and reduced to four regression equations, so we end up with a system like this:

$$INV \cdot 3 = \frac{a1}{a2} \cdot \frac{c2}{c1} \cdot PWR + \frac{a1}{a3} \cdot \frac{c3}{c1} \cdot WCR + \frac{a1}{a4} \cdot \frac{c4}{c1} \cdot TTR \quad (21)$$

$$PWR \cdot 3 = \frac{a2}{a1} \cdot \frac{c1}{c2} \cdot INV + \frac{a2}{a3} \cdot \frac{c3}{c2} \cdot WCR + \frac{a2}{a4} \cdot \frac{c4}{c2} \cdot TTR \quad (22)$$

$$WCR \cdot 3 = \frac{a3}{a1} \cdot \frac{c1}{c3} \cdot INV + \frac{a3}{a2} \cdot \frac{c2}{c3} \cdot PWR + \frac{a3}{a4} \cdot \frac{c4}{c3} \cdot TTR \quad (23)$$

$$TTR \cdot 3 = \frac{a4}{a1} \cdot \frac{c1}{c4} \cdot INV + \frac{a4}{a2} \cdot \frac{c2}{c4} \cdot PWR + \frac{a4}{a3} \cdot \frac{c3}{c4} \cdot WCR \quad (24) .$$

Another set of equations from the theoretical model that can be empirically estimated is the one that corresponds to the system formed by (9), (10), (11) and (12). One simple way to do it is working with the three non-linguistic variables described in section 3, and regressing each partial complexity variable (*INV*, *PWR*, *WCR* and *TTR*) against those categorical variables. What we obtain is something like this:

$$INV = b1r + b1f + b1p \quad ; \quad PWR = b2r + b2f + b2p \quad (25)$$

$$WCR = b3r + b3f + b3p \quad ; \quad TTR = b4r + b4f + b4p \quad (26)$$

where the different *bij* coefficients represent measures of the effect that each category (i.e., each region, family and population size group) has on our partial complexity variables.

If we estimate the system of equations represented in (25) and (26), using ordinary least squares,[4] we obtain a set of coefficients that can be used to build "instrumental variables". Those instrumental variables are created to replace the original partial complexity variables in a new set of regressions, and we will label them as $I\hat{N}V$, $P\hat{W}R$, $W\hat{C}R$ and $T\hat{T}R$. They are formed by the fitted values of the regressions for equations (25)/(26), and their role is to represent the optimal values of *INV*, *PWR*, *WCR* and *TTR* in (21), (22), (23) and (24) (without including any endogenous elements that could make our estimation biased or inconsistent).[5]

The new set of regression equations can therefore be written in the following way:

$$KC = c1 \cdot I\hat{N}V + c2 \cdot P\hat{W}R + c3 \cdot W\hat{C}R + c4 \cdot T\hat{T}R \quad (27)$$

$$INV \cdot 3 = c5 \cdot P\hat{W}R + c6 \cdot W\hat{C}R + c7 \cdot T\hat{T}R \quad (28)$$

$$PWR \cdot 3 = (1/c5) \cdot I\hat{N}V + (c6/c5) \cdot W\hat{C}R + (c7/c5) \cdot T\hat{T}R \quad (29)$$

$$WCR \cdot 3 = (1/c6) \cdot I\hat{N}V + (c5/c6) \cdot P\hat{W}R + (c7/c6) \cdot T\hat{T}R \quad (30)$$

$$TTR \cdot 3 = (1/c7) \cdot I\hat{N}V + (c5/c7) \cdot P\hat{W}R + (c6/c7) \cdot W\hat{C}R \quad (31)$$

where "$c5 = (a1 \cdot c2)/(a2 \cdot c1)$", "$c6 = (a1 \cdot c3)/(a3 \cdot c1)$" and "$c7 = (a1 \cdot c4)/(a4 \cdot c1)$". Their results are reported in Table 3, and were obtained using three-stage least squares.

[4] This estimation was performed using the computing program EViews 3.1. The same software was used for the other regressions whose results are reported in this paper.

[5] For an explanation of the logic behind this procedure, see Kennedy (2008), chapter 10.

Table 3
Three-stage least square regression results

Parameter	Coefficient	Std. Error	t-statistic	Probability
c1	0.001300	0.000799	1.628061	0.1048
c2	0.026399	0.013069	2.019964	0.0445
c3	0.025719	0.005237	4.911356	0.0000
c4	0.392505	0.164358	2.388111	0.0177
c5	9.087457	0.588355	15.445540	0.0000
c6	4.342966	0.223163	19.461000	0.0000
c7	68.872710	3.391096	20.309870	0.0000

With the values that we have found, we can directly compute the parameters of the complexity function (*c1*, *c2*, *c3* and *c4*). Using an indirect calculation, we can also compute values for the parameters of the expressivity function (*a1*, *a2*, *a3* and *a4*). In particular, if we set those values so that they add up to one, we get the coefficients of equation (33), together with the complexity function written as equation (32).

$$C = 0.0013 \cdot INV + 0.026399 \cdot PWR + 0.025719 \cdot WCR + 0.392505 \cdot TTR \quad (32)$$

$$E = 0.0821 \cdot \ln(INV) + 0.1836 \cdot \ln(PWR) + 0.3742 \cdot \ln(WCR) + 0.3601 \cdot \ln(TTR) \quad (33)\ .$$

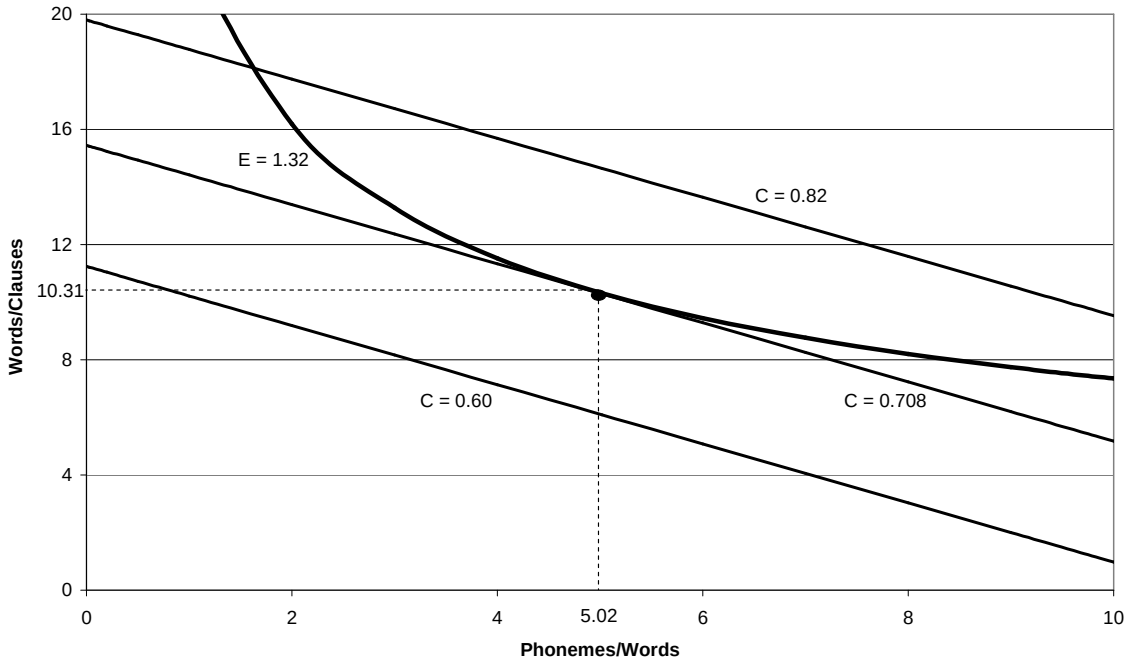

Figure 2. Iso-expressivity and iso-complexity curves

Equations (32) and (33) can be represented in a diagram like the one that appears in Figure 2, in which we have drawn one particular case of "E" (the one that corresponds to the expressivity levels implied by the average values of *INV*, *PWR*, *WCR* and *TTR*) and three particular cases of "C" (then one in which that function equates the average level of *KC*, plus two

additional ones). As we can see, this diagram is depicted in the space of *PWR* vs. *WCR*, and in it we find that our iso-expressivity curve is tangent to the iso-complexity line for which *C = 0.708* (which is the average level of *KC* in our sample). This means that such level of global complexity is the minimum one that could be obtained if we require that a language has the expressivity implied by the average levels of *INV*, *PWR*, *WCR* and *TTR*. Moreover, in that tangency point we see that the values for *PWR* and *WCR* are the ones that correspond to the average values of those variables (i.e., *PWR = 5.02* and *WCR = 10.31*).

If we make a small variation in our model, we can also use it to estimate partial correlation coefficients between the complexity variables. In order to do that, we have to write equations (28)/(31) in the following way:

$$INV \cdot 3 = c5 \cdot P\hat{W}R + c6 \cdot W\hat{C}R + c7 \cdot T\hat{T}R \quad (34)$$

$$PWR \cdot 3 = c8 \cdot I\hat{N}V + c6 \cdot c8 \cdot W\hat{C}R + c7 \cdot c8 \cdot T\hat{T}R \quad (35)$$

$$WCR \cdot 3 = c9 \cdot I\hat{N}V + c5 \cdot c9 \cdot P\hat{W}R + c7 \cdot c9 \cdot T\hat{T}R \quad (36)$$

$$TTR \cdot 3 = c10 \cdot I\hat{N}V + c5 \cdot c10 \cdot P\hat{W}R + c6 \cdot c10 \cdot W\hat{C}R \quad (37)$$

where we assume that *c8*, *c9* and *c10* are not necessarily equal to *1/c5*, *1/c6* and *1/c7*. Let us now define the correlation coefficients between our variables using the following formula:

$$r_{xy} = -\sqrt{\frac{c_{xy} \cdot c_{yx}}{9}} \quad (38)$$

where r_{xy} is the partial correlation coefficient between variables *x* and *y*, c_{xy} is the regression coefficient that corresponds to $\hat{y}$ in the equation where the dependent variable is *x·3*, and c_{yx} is the regression coefficient that corresponds to $\hat{x}$ in the equation where the dependent variable is *y·3*.[6]

Table 4
Partial correlation coefficients between complexity variables

Complexity variable	INV	PWR	WCR	TTR
Phonological inventory	1.0000			
Phoneme/word ratio	-0.2322	1.0000		
Word/clause ratio	-0.3720	-0.2592	1.0000	
Type/token ratio	-0.3947	-0.2750	-0.4406	1.0000

Using the results obtained in our new set of regressions, we used equation (38) to calculate the coefficients reported in Table 4. All of them turned out to be statistically significant at a 1% probability level, and the largest absolute value is the one that corresponds to the relationship between *WCR* and *TTR*. Note also that some variables that display positive product-moment correlation coefficients in Table 2 (*INV* vs. *WCR*, and *PWR* vs. *TTR*) are now negatively related. This is consistent with the idea that partial complexity measures are linked through the interaction between functions "C" and "E", and must therefore be negatively correlated in all cases.

[6] For an explanation of the logic behind this formula, see Prokhorov (2002).

5. Concluding remarks

The two points that we believe are more important in this paper are related to the use of the proposed optimization model, and to its implementation through the concept of Kolmogorov complexity. On one hand, we think that our model is an elegant theoretical approach to the general problem of language as a self-regulated system, and that it is also good to include the interaction that the elements from that system may have with external forces such as phylogenetic, geographic and demographic factors.

On the other hand, we see the concept of Kolmogorov complexity (and its approximation through the ratio between the sizes of a compressed file and an original text file) as a promising empirical approach to global language complexity. Due to the fact that it is a measure that can be applied to different texts, it can also be correlated to other (partial) complexity measures for those texts, which can in turn be seen as their internal determinants.

The logic behind our results is that the relationship between the different complexity measures can be interpreted as the outcome of a process in which the language system defines certain levels of partial complexity in order to minimize a global complexity function, subject to an expressivity constraint. Using particular functional forms for those relationships, we were able to illustrate them through various parameters that are estimated in a simultaneous-equation regression procedure. In that procedure, we also used information from non-linguistic variables that define each observation in our sample (i.e., the region and family to which each language belongs, and its size in terms of number of speakers).

However, the empirical illustration included in this paper is not intended to test the accuracy of the proposed model to fit actual data. Its purpose is to show how the theoretical variables of the model can be interpreted as observable variables, and how those observable variables can be used to figure out plausible "shapes" for the functions postulated in the theoretical model. Of course, the model could also be used, in a different setting, to be contrasted with another theoretical alternative that provides a different explanation for language complexity phenomena.

Another possible use of the optimization model developed in this paper has to do with testing different definitions for the global complexity variables (apart from Kolmogorov complexity). It could also be possible to use the empirical version of the model to test different functional forms for the complexity and expressivity functions, since our linear and logarithmic versions of those functions are just one, relatively simple, alternative to write the relationships embedded in the theoretical model. That alternative can certainly be contrasted with other additional specifications.

Finally, the model could be applied in different contexts that were not necessarily cross-linguistic. An alternative sample to the one used could consist of texts written in the same language, but belonging to different authors, or genres, or styles, or time periods that cover different stages in the evolution of language.

Acknowledgments

This paper was financed by the Research Fund of CEMA University (Buenos Aires, Argentina), as part of the project entitled "The Use of Econometric Methods to Analyze Linguistic Problems". I thank Gabriel Altmann, Damián Blasi, Mariana Conte Grand and Ezequiel Koile for their useful comments.

References

Atkinson, Quentin (2011). Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa. *Science 332, 346-349*.

Balasubrahmanyan, Viddhachalam & Sundaresan Naranan (2002). Algorithmic Information, Complexity and Zipf's Law. *Glottometrics 4, 1–26.*

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 59(1), 1-26.*

Chiang, Alpha & Kevin Wainwright (2005). *Fundamental Methods of Mathematical Economics*, 4th edition. Boston: McGraw-Hill.

Coloma, Germán (2015). The Menzerath-Altmann Law in a Cross-Linguistic Context. *SKY Journal of Linguistics 28, 139-159.*

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

Culicover, Peter (2013). *Grammar and Complexity*. Oxford: Oxford University Press.

Ehret, Katharina & Benedikt Szmrecsányi (2015). An Information-Theoretic Approach to Assess Linguistic Complexity". In: R. Baechler & G. Seiler (eds.): *Complexity, Variation and Isolation*. Berlin: De Gruyter.

Fenk-Oczlon, Gertraud & August Fenk (2005). Crosslinguistic Correlations Between Size of Syllables, Number of Cases, and Adposition Order. In: G. Fenk-Oczlon & C. Winkler (eds.): *Sprache und Natürlichkeit: 75-86*. Tübingen: Narr.

Ferrer-i-Cancho, Ramón (2014). *Optimization Models of Natural Communication*, mimeo. Barcelona: Universitat Politècnica de Catalunya.

Futrell, Richard, Kyle Mahowald & Edward Gibson (2015). Large-Scale Evidence of Dependency Length Minimization in 37 Languages. *PNAS 112(33), 10336-10341.*

Hawkins, John (2004). *Efficiency and Complexity in Grammars*. New York: Oxford University Press.

Hay, Jennifer & Laurie Bauer (2007). Phoneme Inventory Size and Population Size. *Language 83, 388-400.*

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

Juola, Patrick (2008). Assessing Linguistic Complexity. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 89-108.* Amsterdam: John Benjamins.

Ke, Jinyun, Mieko Ogura & William Wang (2003). Optimization Models of Sound Systems Using Genetic Algorithms. *Computational Linguistics 29, 1-18.*

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

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

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

Kolmogorov, Andrei (1963). On Tables of Random Numbers. *Sankhya 25, 369-376.*

Kusters, Wouter (2003). *Linguistic Complexity*. Utrecht: LOT.

Li, Ming & Paul Vitányi (1997). *An Introduction to Kolmogorov Complexity and its Applications*, 2nd edition. New York: Springer.

McWhorter, John (2001). The World's Simplest Grammars Are Creole Grammars. *Linguistic Typology 5, 125-166*.

Miestamo, Matti (2008). Grammatical Complexity in a Cross-Linguistic Perspective. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 23-41*. Amsterdam: John Benjamins.

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

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 10, 1-40*.

Sinnemäki, Kaius (2008). Complexity Trade-Offs in Core Argument Marking. In: M. Miestamo, K. Sinnemäki & F. Karlsson (eds.), *Language Complexity: Typology, Contact and Change: 67-88.* Amsterdam: John Benjamins.

Sundaram, Rangarajan (1996). *A First Course in Optimization Theory*. New York, Cambridge University Press.

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

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

Appendix
Dataset from "The North Wind and the Sun"

Language	Family	Region	Size	INV	PWR	WCR	TTR	KC
Amharic	AA	NAfrica	Medium	41	6.958	11.88	0.7263	0.6239
Apache	Other	NAmerica	Small	57	4.907	7.87	0.6017	0.6167
Arabic	AA	WAsia	Large	35	5.741	9.44	0.7647	0.6962
Arrernte	Other	Pacific	Small	35	5.892	6.17	0.6351	0.6150
Basque	Other	SEurope	Small	33	4.831	11.86	0.6506	0.7690
Bemba	NC	SAfrica	Medium	46	5.506	9.88	0.7468	0.7138
Bengali	IE	CAsia	Large	36	4.371	10.50	0.7143	0.7154
Berber	AA	NAfrica	Medium	37	3.873	8.78	0.7468	0.8087
Burmese	ST	EAsia	Medium	70	7.143	6.00	0.9048	0.9195
Cantonese	ST	EAsia	Medium	85	3.857	9.10	0.6484	0.7739
Chickasaw	Other	NAmerica	Small	34	8.316	5.70	0.6667	0.6376
Dinka	NS	NAfrica	Small	48	4.000	13.70	0.5474	0.7030
English	IE	NEurope	Large	35	3.389	12.56	0.5575	0.6945
French	IE	SEurope	Large	33	3.176	12.00	0.5926	0.7205
Georgian	Other	WAsia	Medium	38	6.058	7.67	0.8116	0.7731
German	IE	NEurope	Large	53	4.147	10.90	0.6514	0.6972
Greek	IE	SEurope	Medium	28	4.165	12.78	0.5478	0.7046
Hausa	AA	SAfrica	Medium	48	3.904	13.83	0.5241	0.6094
Hebrew	AA	WAsia	Medium	35	5.910	8.09	0.8202	0.7400
Hindi	IE	CAsia	Large	45	3.766	15.50	0.6290	0.6252
Hungarian	Other	NEurope	Medium	39	4.310	10.00	0.6300	0.7418
Igbo	NC	SAfrica	Medium	50	3.358	13.25	0.5094	0.8044
Irish	IE	NEurope	Small	46	3.147	18.43	0.5969	0.7421
Japanese	Alt	Pacific	Large	26	5.045	9.78	0.6023	0.7145
Kabiye	NC	SAfrica	Medium	39	4.758	10.11	0.6923	0.7441
Korean	Alt	EAsia	Medium	37	6.350	8.57	0.7833	0.8536
Malay	Aus	Pacific	Large	24	6.167	9.75	0.6154	0.6485
Mandarin	ST	EAsia	Large	43	4.385	9.60	0.6875	0.7940
Mapudungun	Other	SAmerica	Small	28	4.800	8.33	0.4800	0.7644
Nara	NS	NAfrica	Small	45	4.315	9.82	0.5463	0.6555
Nepali	IE	CAsia	Medium	38	5.340	10.44	0.8085	0.7198
Persian	IE	WAsia	Medium	35	5.308	10.11	0.7143	0.7220
Portuguese	IE	SEurope	Large	45	3.878	12.25	0.6429	0.7747
Quichua	Other	SAmerica	Medium	26	6.589	8.18	0.7556	0.6037
Russian	IE	NEurope	Large	48	4.825	10.78	0.7320	0.7261
Sahaptin	Other	NAmerica	Small	46	6.579	7.13	0.6140	0.7923
Sandawe	Other	SAfrica	Small	74	5.716	7.44	0.7612	0.6993
Seri	Other	NAmerica	Small	26	3.777	14.27	0.4459	0.5142
Shiwilu	Amaz	SAmerica	Small	25	7.750	7.71	0.6759	0.5322
Spanish	IE	SEurope	Large	29	4.381	10.78	0.6186	0.7761
Tajik	IE	WAsia	Medium	28	5.477	12.57	0.7159	0.6559
Tamil	Other	CAsia	Medium	25	6.763	8.89	0.6750	0.5686
Tausug	Aus	Pacific	Small	20	5.018	9.50	0.4825	0.6019
Temne	NC	SAfrica	Medium	37	3.568	11.36	0.5440	0.6302
Thai	Other	EAsia	Medium	66	3.664	11.91	0.5649	0.6437
Trique	OM	NAmerica	Small	101	3.355	10.70	0.5794	0.7059
Turkish	Alt	WAsia	Medium	30	6.631	7.22	0.7538	0.8629
Vietnamese	Other	EAsia	Medium	110	2.855	16.71	0.5299	0.7622
Yine	Amaz	SAmerica	Small	21	9.164	6.10	0.7049	0.6426
Zapotec	OM	NAmerica	Medium	35	3.759	9.67	0.6552	0.8440