Language Sciences 59 (2017) 148–158

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

# Complexity trade-offs in the 100-language WALS sample

Germán Coloma

*CEMA University, Av. Cordoba 374, Buenos Aires, C1054AAP, Argentina*

ARTICLE INFO

*Article history:*
Received 19 March 2016
Received in revised form 22 September 2016
Accepted 15 October 2016
Available online 15 November 2016

*Keywords:*
Complexity trade-off
WALS
Binary variables
Partial correlation

ABSTRACT

In this paper we use data from the World Atlas of Language Structures (WALS) for a balanced sample of 100 languages and 60 different features. The values for all those features are interpreted as binary complexity variables, which are subject to statistical correlation analyses (looking for the possible existence of complexity trade-offs). To do that we use standard correlation coefficients but also partial correlation coefficients, which control for the effect of other linguistic and non-linguistic factors (geographic location, genetic affiliation, population size). We end up with the conclusion that several important complexity trade-offs exist, but they tend to be hidden by other elements. Their most evident signals are the facts that negative correlations between complexity variables increase when we control for other factors, and that any language is more complex than any other language in the sample in at least one feature.



## 1. Introduction

The World Atlas of Language Structures (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.

None of the languages included in WALS has data for all the reported features, and no feature reports values for all languages, either. The language with the maximum number of reported features is English (159 entries), but there are many cases such as Asturian (Indo-European, spoken in Spain) or Taiwanese (Sino-Tibetan, spoken in Taiwan) that have information for only one feature. The feature with the greatest number of data points (83A: Order of object and verb) displays 1519 entries, while the feature with the fewest data points (141A: Writing systems) has information for only 6 languages.

The editors of WALS have selected a sample of 100 languages that the authors of the different chapters of the atlas are supposed to include in their reports "if at all possible". Those languages form a relatively balanced sample of genealogical and areal diversity, although other considerations were also taken into account in the selection, related to the availability of detailed grammatical descriptions and the overall importance of the languages.[1]

Making use of the fact that WALS concentrates a very large amount of structural data, and that it has privileged the compilation of information about a particular sample of languages, in this paper we use the data from that sample to conduct a series of statistical analyses aimed at the detection of possible relationships between the values of 60 different features

*E-mail address:* gcoloma@cema.edu.ar.

[1] The complete list of languages is reproduced in Appendix 1.

http://dx.doi.org/10.1016/j.langsci.2016.10.006
0388-0001/

included in WALS. The selected features are from different language domains, but in all cases they have been coded as variables that can be interpreted as measures of complexity.[2]

With the variables chosen, we proceeded to calculate correlations, using both standard (Pearson) coefficients and partial coefficients. When those coefficients are negative, we interpret them as a sign of the possible existence of complexity trade-offs. A complexity trade-off is a situation in which a higher level of complexity for a certain language component appears with a lower level of complexity for another component. This kind of relationship can occur between features that belong to the same language domain (e.g., phonology, morphology, syntax), or between features that belong to different domains.

The rest of the paper is organized as follows. In Section 2 we review the literature about complexity trade-offs, especially the research that has looked for negative correlations between different language variables. In Section 3 we describe the dataset that we have assembled. In Section 4 we analyze the main statistics calculated for that dataset, in terms of correlations between the variables. Finally, in Section 5 we make some concluding remarks.

## 2. Review of the literature

The literature on complexity trade-offs in cross-linguistic contexts is relatively large and diverse. This diversity arises at the level of the definition of complexity, at the way in which that complexity is measured, and also in the results obtained by the different authors. One of the most cited papers concerning this issue is McWhorter (2001), which proposed the design of complexity metrics for the different subsystems of language (e.g., phonology, morphology, syntax) based on the idea that one language is more complex than another one if it encompasses "more overt distinctions and/or rules".[3] Applying this definition, McWhorter concluded that creole languages (i.e., natural languages that have developed as a simplified version of another natural language) are simpler than non-creole languages, since they lack a number of overspecifications that older languages typically have.

McWhorter's hypothesis went against a long tradition in linguistics that considered that all languages were equally complex.[4] That tradition, however, was not based on the systematic measurement of complexity indicators but on different theoretical approaches that assumed equal complexity. However, since the quantitative linguistics' literature began to measure complexity and to analyze the possible correlation between complexity measures, that hypothesis has generally been shown to be false.

One of the most important papers in this line of research is Shosted (2006). In that study, the author sought the possible existence of a negative correlation between phonological and morphological complexity, using a sample of 32 languages. He measured phonological complexity using the theoretical number of possible syllable types in each language, and morphological complexity using the number of inflectional categories of the verbs, and concluded that there was no statistically significant correlation between those two concepts.

In the same line of research, Nichols (2009) measured cross-linguistic complexity in five language domains (phonology, inflectional synthesis, classification, syntax and lexicon) using different indices, calculated as averages of several complexity measures within a sample of 68 languages. She found no significant negative correlation between any of those indices, but she did find a significant positive correlation between her measure of synthesis (morphological) complexity and her measure of syntactic complexity.

While both Shosted and Nichols measured linguistic complexity using typological (or theoretical) measures, there is a different approach in the literature that uses empirical measures. Fenk-Oczlon and Fenk (1999), for example, calculated several linguistic ratios (phonemes per syllable, syllables per word, phonemes per word, words per clause, etc.) that came from translations of a number of simple German sentences into 33 different languages. Later on they computed correlation coefficients between those ratios, and they found negative correlations between many of the analyzed variables.[5]

Another branch of the literature is the one that relates word length (measured using a sample of words from different languages) to phonological complexity (measured using theoretical measures such as phoneme inventory size or number of tones). The first paper in this domain is Nettle (1995), which found a strong negative correlation between phoneme inventory size and phonemes per word in a sample of ten languages. This line of research was later followed up by studies such as Wichmann et al. (2011), Moran and Blasi (2014), which used much larger datasets.

Another example of a study that looks for possible phonological complexity trade-offs is Maddieson (2007), that measured that phenomenon using three different typological indicators: inventory size (of both consonants and vowels), tone systems, and elaboration of the syllable inventory. That paper also looked for correlations between those measures, but it only found a significant positive correlation between consonant inventory and syllable structure, and a (less significant) negative correlation between syllable structure and tone complexity.

Another example of a quest for complexity trade-offs within a particular language sub-system is Sinnemäki (2008), which focused on three alternative measures of (syntactic) complexity for core argument marking, which are head marking,

[2] The complete list of variables is reproduced in Appendix 2.

[3] This definition is related to what some authors call "absolute complexity", which can be defined as "the number of parts in a system". On this topic, see Miestamo (2008).

[4] For a good summary of this literature, see Sampson (2009), Joseph and Newmeyer (2012).

[5] On this point, see also Fenk-Oczlon and Fenk (2008), Coloma (2016).

dependent marking and rigid word order.[6] After defining those measures based on their capacity to discriminate core arguments, Sinnemäki found a significant negative correlation between dependent marking and rigid word order, and nonsignificant negative correlations between head marking and dependent marking, and between head marking and rigid word order.

Some examples in the literature on linguistic complexity trade-offs have used data from WALS. This has occurred with the already cited papers by Maddieson (2007), Sinnemäki (2008), Moran and Blasi (2014), and with numerous other articles concerning linguistic complexity and correlation, such as Parkvall (2008), Dahl (2011) and Matasovic (2014).[7]

As a general assessment related to the literature on complexity trade-offs in cross-linguistic contexts, we can say that those trade-offs seem to be more prevalent when authors use empirical measures and rarer when they use typological measures. They also seem to be more common when we look at studies about a single language domain (e.g., phonology, syntax), and less common when the comparison is between measures that come from different language domains.

In this paper, we look for complexity trade-offs within and between different language domains. This is because our dataset comprises a relatively large number of linguistic features, which belong to different categories. All of our complexity measures, however, are theoretical or typological in nature, since they do not come from specific texts but from the languages' grammars, compiled by the authors of the different WALS chapters.

## 3. Description of the data

As we mentioned in the introduction, our dataset contains information for 60 variables from the languages included in the 100-language WALS sample. Those variables were created using information about specific features provided by WALS, and they are all "binary variables", i.e., variables that take a value equal to one in certain cases, and zero otherwise.

The criterion used to define the variables is to assign a value equal to one to the languages that are more complex in terms of a certain linguistic feature, and a value equal to zero to the languages that are simpler. To do that, we selected all the features for which the information that appears in WALS allows us to classify languages according to some kind of complexity metric. For example, feature 1A (consonant inventories) is the basis of a variable for which "complexity" means that the number of consonant phonemes is large or moderately large (26 or more phonemes), while "simplicity" means that the number of those phonemes is smaller (25 or fewer).[8] Under the same line of reasoning, the variable based on feature 66A (past tense) assigns a value equal to one (complex) to languages that make some kind of grammatical distinction between past and non-past verbal forms, and a value equal to zero (simple) to languages that do not make any grammatical distinction between those forms.

This way of defining language complexity is of course not perfect, since it does not take into account some aspects related to the number of component types, the number of possible interactions between the components, or the number of rules determining those interactions. It nevertheless has the advantage of being straightforward to operationalize, in a way that seems tractable for use in a sample that includes a lot of languages that are very different from each other.

Many features reported in WALS are not included in the analysis, mainly because they do not provide a basis for generating variables that can easily be interpreted as measures of complexity. Feature 3A (consonant-vowel ratio) is an example of that, since it is not clear that a language having a large ratio between consonants and vowels is more complex than a language that has fewer consonants per vowel. The same occurs with feature 31A (gender systems), since having a sex-based gender system (i.e., a system in which nouns are either masculine or feminine) is neither simpler nor more complex than having a non-sex-based gender system. This contrasts with the use that we make of feature 30A (number of genders), which generates a variable under which a language is complex if its nouns are marked by some kind of gender distinction (sex-based or non-sex-based), and simple if it has no grammatical gender distinctions.

The construction of our dataset was also conditioned by the fact that there are many observations for which WALS does not provide complete information. This is because only five languages in the sample (English, Finnish, French, Indonesian and Turkish) have data points for all the features that we selected, and only seven of the selected features (48A, 100A, 104A, 107A, 108A, 109A, and 113A) have data points for all the languages included in the sample. The number of missing data points is equal to 532, which is a relatively small figure if we consider the total number of points in the dataset (which is equal to 6000). Nevertheless, the reconstruction of those missing points required a considerable effort, and several strategies were used for that task.

Many missing data points were inferred by looking at other data points that belong to dialects of the same language (e.g., Gulf Arabic instead of Egyptian Arabic, or Huallaga Quechua instead of Imbabura Quechua). In other cases for which that procedure was not possible, we looked for a language that belongs to the same genus as the missing data point language (e.g., Cebuano instead of Tagalog, or Koyra Chiini instead of Koyraboro Senni). In some other cases we had to do a more indirect

[6] See also Sinnemäki (2014).

[7] Gordon (2016) is also worth noting at this point, because in his analyses concerning different phonological features he always uses information from the 100-language WALS sample.

[8] Note that some variables come from information that is originally represented by numerical values. In WALS, however, all the information is reported as belonging to different categories (e.g., large, moderately large, moderately small, etc.), and not as numerical variables. That is why we have preferred to transform all the features reported in WALS into categorical binary variables, and then proceed with our numerical analysis treating all the variables in the same way.

reconstruction, using data from languages that are not so close genetically but are linked through some areal relationships (e.g., Yurok instead of Karok, or Igbo instead of Yoruba). Those procedures, of course, have a certain degree of imprecision and are subject to measurement errors, but we nevertheless believe that the probability of those errors is considerably reduced by the fact that we classify languages in binary groups (simple or complex), and those groups tend to be distributed in a relatively regular fashion across language families and regions.

The 60 features that were used to build our 100-language dataset belong to several WALS areas, which are: phonology (11 features), morphology (6 features), nominal categories (14 features), nominal syntax (4 features), verbal categories (10 features), word order (5 features), simple clauses (8 features), and lexicon (2 features). When the information that appears in WALS for a particular language feature permits dividing the sample in different ways, we make the division between simple and complex languages so that the number of data points in each set is relatively even. For example, feature 41A (distance contrasts in demonstratives) classifies languages in five groups, according to the number of distance contrasts that are marked by different words or affixes. Based on their distribution in those groups, we consider that languages with no contrasts and languages with only two-way distance contrasts (e.g., English, which only contrasts "this" and "that") are simple, and languages with three-way or more contrasts (e.g., Spanish, which contrasts "este", "ese" and "aquel") are complex. With that division, 48 languages are considered to be complex and 52 languages are considered to be simple, while if we had used an alternative division we would have ended up with 96 complex languages and 4 simple ones (if we had used the label "simple" only for languages with no distance contrasts), or with 91 simple languages and 9 complex ones (if we had considered three-way contrasts as "simple").

In several circumstances, the variables created to measure language complexity are subject to different possible interpretations concerning which feature values are seen as "complex" and which ones are seen as "simple". In those cases, we always used an interpretation under which complexity means "more overt distinctions and/or rules" (McWhorter, 2001). This implies, for example, that a language that has definite articles (feature 37A) is considered to be complex, while another language that does not have those articles is considered to be simple in that language dimension. Of course, the last language may have another means to express definiteness which does not imply the use of articles, but concerning that particular feature the first language is more complex than the second one because it has an overt rule governing the use of definite articles that the second language lacks.

Taking into account the 100 languages and the 60 variables as a whole, we end up with 2637 data points (44%) that represent complexity and 3363 data points (56%) that represent simplicity. The variable with the largest fraction of complex points is the one that corresponds to feature 130A (finger and hand), with 91 languages that use different words for "finger" and "hand", and only 9 languages that use the same word. The variable with the smallest fraction of complex points is the one that corresponds to feature 11A (front rounded vowels), with 7 languages that possess at least one front rounded vowel as a distinct phoneme, and 93 languages that lack those vowels. If we look at languages, the largest fraction of complex data points corresponds to Abkhaz, with 41 complex features and 19 simple ones, while the languages with the largest fraction of simple data points are Hmong Njua and Thai, with 14 complex features and 46 simple ones.

The fact that some languages have a lot of complex features while others have very few does not mean that the "most complex" languages are more complex than the simplest ones in all possible dimensions. Quite the contrary, in the 100-language WALS sample we have found no language that is so simple that it does not display a higher complexity value than any of the other languages in at least one particular feature. Indeed, if we check all the 4950 possible language pairs that can be compared within the sample, we find that in all of them there is always at least one feature for which the first language is more complex than the second one, and at least another feature for which the second language is more complex than the first one.

Consider, for example, one of the most extreme cases, which is the comparison between Zulu and Swahili. Zulu has 32 complex features, which is well above the mean (it ranks 14th in the whole sample), while Swahili has only 26 complex features, slightly below the mean (which is equal to 26.37). Zulu has higher complexity values than Swahili for 7 different features (7A: glottalized consonants, 13A: tone, 41A: distance contrasts in demonstratives, 64A: nominal and verbal conjunction, 65A: perfective/imperfective aspect, 129A: hand and arm, and 143F: postverbal negative morphemes), but Swahili is more complex than Zulu in one feature (37A: definite articles). Therefore, if we consider Swahili's having definite articles while Zulu does not to compensate for the other phonological, syntactic, lexical and verbal features in which Zulu is more complex than Swahili, then we can say that, in a certain sense, Swahili and Zulu are equally complex (or at least, that neither of them is more complex than the other in all the possible dimensions of language analysis).[9]

## 4. Correlation analysis

Having selected 60 complexity variables for the present study, we next computed correlation coefficients to look for complexity trade-offs among those variables. Those correlations come from relating columns of 100 numbers each. As all our

[9] Swahili and Zulu are also languages that share a large number of identical features, since they have the same value for 52 out of the 60 binary variables that we have defined for the 100-language WALS sample. The case with the largest overlapping is the one of Vietnamese and Thai, which have the same value for 57 variables. These two languages only differ in features 7A and 37A (for which Vietnamese is more complex) and feature 129A (for which Thai is more complex).

**Table 1**
Contingency table between syllable structure and tone.

| Syllable structure/tone | No tone | Tonal | Total |
|---|---|---|---|
| Simple | 39 | 29 | 68 |
| Complex | 29 | 3 | 32 |
| Total | 68 | 32 | 100 |

variables are binary, those numbers are alternatively equal to zero or one, and can be represented in contingency tables like Table 1. That table exemplifies the case of the variables corresponding to features 12A (syllable structure) and 13A (tone).

Looking at the figures reported in Table 1, we can see the number of observations that belong to the four possible cases in the sample: no-tone languages with a simple syllable structure (39 observations), no-tone languages with a complex syllable structure (29 observations), tonal languages with a simple syllable structure (29 observations), and tonal languages with a complex syllable structure (3 observations). Those figures show that the most popular pairing is the one that corresponds to languages that are simple in the two dimensions that we are contrasting, and that the least popular is the one in which languages are complex in both tone and syllable structure.

But Table 1 can also be used to calculate the standard (Pearson) correlation coefficient between tone and syllable structure in a compact way. Indeed, using the figures that appear in our contingency table, this correlation coefficient (r) can be computed using the following formula:

$$r = \frac{N(S,NT)\cdot N(C,T) - N(S,T)\cdot N(C,NT)}{\sqrt{N(S)\cdot N(C)\cdot N(NT)\cdot N(T)}} = \frac{39\cdot 3 - 29\cdot 29}{\sqrt{68\cdot 32\cdot 68\cdot 32}} = -0.3327;$$

where $N(S,NT)$, $N(C,T)$, $N(S,T)$ and $N(C,NT)$ are the numbers in each of the four cells of the contingency table, and $N(S)$, $N(C)$, $N(NT)$ and $N(T)$ are the total numbers of observations that imply simple syllable structures, complex syllable structures, no-tone languages and tonal languages.[10]

In this case, the outcome of this formula indicates the existence of a negative correlation, which turns out to be statistically significant.[11] It implies that languages that have a relatively simple syllable structure are more likely to use tone as a feature to distinguish individual words or the grammatical forms of words, while languages with a more complex syllable structure do not generally use tone as a distinctive feature. This can be seen as an example of a complexity trade-off, since complexity in one dimension (syllable structure) is related to simplicity in another dimension (tone) and vice versa.

If we compute the standard correlation coefficients for the 1770 pairings of the 60 variables (i.e., 60 times 59, divided by 2), we find that 795 of them (44.9%) are negative and 975 (55.1%) are positive.[12] Most calculated coefficients, however, are statistically nonsignificant, and only 85 of them (4.8%) are at the same time negative and significant at a 5% probability level (i.e., negative and greater than 0.2 in absolute value). We have also found 186 coefficients (10.5%) that are simultaneously positive and significant.

Among the negative and significant coefficients, we can mention some cases such as the one corresponding to features 100A (verbal person marking) and 119A (nominal and locational predication), whose correlation coefficient ($r = -0.3021$) indicates that languages which mark the person that is implied by a verbal form are more likely to use the same verb for nominal predication (e.g., John is a man) and locational predication (e.g., John is in America). Another clear case of complexity trade-off between two variables is the one that appears between features 129A (hand and arm) and 130A (finger and hand). The negative correlation coefficient between these lexical variables (which is equal to −0.2359) indicates that, when a language uses the same word for “hand” and “arm”, then it is likely to use a different word for “finger”.

The fact that we have found many nonsignificant coefficients (and more positive than negative coefficients) can be interpreted in different ways. On one hand, we can see it as an indication that complexity trade-offs are not very important in the 100-language WALS sample. On the other hand, we can also see it as a signal of the existence of methodological problems, which are not letting us to adequately capture some trade-offs that can exist but are actually hidden by those problems.

The use of standard correlation coefficients to detect the existence of complexity trade-offs between different language features has the advantage that it is relatively easy to compute, but it has the disadvantage that it does not take into account the possible interaction with other relevant variables.[13] To solve that problem, we can calculate partial correlation coefficients, which are measures of the linear dependence for a pair of variables when the influence of other variables is removed. To

[10] This type of correlation coefficient, which is suitable for binary variables, is also known as the “phi coefficient” or the “mean square contingency coefficient”. For a more complete explanation of its properties, see Gingrich (2004), chapter 11.

[11] Statistical significance means that it is not likely that the true value of a coefficient is in fact zero. In this particular case, the correlation coefficient is −0.3327, and the associated probability that this value is not different from zero is 0.0007. 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 0.05) if they are greater than 0.2 in absolute value.

[12] Because we are computing so many correlation coefficients, some will be spurious. At the probability level of 0.05 that we have chosen, approximately five in every 100 tests (about 88 in all) may have spurious significance. But as we obtained many more significant outcomes than this (85 negative and 186 positive, which make a total of 271 coefficients), we can conclude that most significant outcomes are real.

[13] For a good explanation of this disadvantage, in the context of typological correlations calculated using WALS, see Dryer (2009).

compute those coefficients, it is necessary to use some additional procedures. One possibility is to begin with a complete correlation matrix for all the variables under analysis, and then invert that matrix. Once we do that, we can use the following formula:

$$r = -\frac{p_{xy}}{\sqrt{p_{xx} \cdot p_{yy}}};$$

where $p_{xy}$ is the coefficient that corresponds to the pair of variables $x$ and $y$ in the inverted correlation matrix, and $p_{xx}$ and $p_{yy}$ are the coefficients that correspond to those variables in the main diagonal of the same inverted correlation matrix.

This process of matrix inversion is actually one of the possibilities that we can use to obtain partial correlation coefficients. Another one is to run a set of 60 different regressions, in which each variable is regressed against the other 59 ones. Both procedures have the same goal, which is pulling out the effects that the remaining 58 variables may have on each pair of variables that we are interested in.[14]

If we perform that procedure in our 100-language sample with 60 variables, we end up with another matrix whose 1770 correlation coefficients for all the possible pairs of variables are either negative (878 cases, 49.6%) or positive (892 cases, 50.4%). The number of statistically significant negative coefficients is now greater than that obtained for the standard correlation matrix, since in this partial correlation matrix there are 162 negative coefficients (9.15%) that are greater than 0.2 in absolute value. Among them, we can mention the case of features 1A (consonant inventory) and 2A (vowel quality inventory), whose negative correlation coefficient ($r = -0.3163$) indicates that languages with more consonants tend to have fewer phonemically distinct vowel qualities and vice versa. Computed as a standard coefficient, this correlation was negative but failed to be statistically significant (since it was equal to $-0.1622$).[15]

Another coefficient that can be interpreted as indicative of a complexity trade-off and is statistically significant if we compute it as a partial correlation (and not if we calculate it as a standard correlation) is the one that corresponds to features 23A (locus of marking) and 30A (number of genders). This correlation coefficient (equal to $-0.3437$ using partial correlation techniques, and equal to $-0.0750$ if we compute it as a standard correlation) indicates that languages which do not mark the direct object with a particular affix are more likely to possess grammatical gender distinctions than languages that do mark the direct object overtly. The same situation occurs for correlations such as those between features 68A (perfect tense) and 70A (imperative), between features 84A (order of object, oblique and verb) and 92A (polar question particles), etc.

One additional instrument that we can include to improve the measurement of possible complexity trade-offs between the different features reported in the 100-language WALS sample, is to use other variables that are not linguistic in nature. These variables may actually be correlated with some of the complexity measures that we have included in our sample, and this correlation may interfere with the measurement of possible trade-offs between our variables. Once we include a set of nonlinguistic variables into the pool of factors that we use to calculate partial correlation coefficients, that interference is eliminated (in the same way that we have already removed the influence of the other linguistic factors when we compute the pairwise correlation coefficient between two particular complexity variables).

One relatively straightforward possibility for dealing with nonlinguistic factors is to add variables that represent areal features, such as binary variables that take a value equal to one when a language belongs to a certain region, and zero otherwise. As WALS divides languages in six macro-areas, we have created five variables that correspond to languages located in those areas, which are Africa (16 languages), Australia (7 languages), Papunesia (17 languages), North America (18 languages) and South America (13 languages).[16]

Three additional binary variables were created to deal with the possible existence of some genetically-driven forces. These correspond to languages that belong to three families with a relatively large number of observations in the 100-language WALS sample, which are Indo-European (8 languages), Austronesian (8 languages) and Niger-Congo (7 languages). We also included three other binary variables that take a value equal to one when a language belongs to one of the following regions within a particular macro-area: South East Asian languages (Burmese, Hmong Njua, Mandarin, Meithei, Thai and Vietnamese), Mesoamerican languages (Jakaltek, Mixtec, Otomi, Rama, Yaqui and Zoque), and Amazonian languages (Apurina, Barasano, Canela-Kraho, Hixkaryana, Piraha, Sanuma, Wari and Yagua).

The last non-linguistic binary variable that we have used has to do with some differences reported in the literature between languages spoken by a large number of speakers (major languages) and languages spoken by a smaller number.[17] This variable assigns a value equal to one to the 33 languages whose total number of speakers is above 5 million people (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), and a value equal to zero to the remaining 67 languages.

Using the 12 newly included binary variables, we have calculated new partial correlation coefficients. Applying the same procedure described above, we ended up with coefficients that signal 881 negative relationships (49.8%) and 889 positive

[14] For a more complete explanation of the concept or partial correlation, see Prokhorov (2002).

[15] Note that this result is different from the one obtained by Maddieson (2007), who found a positive but nonsignificant (standard) correlation between consonant inventory and vowel inventory, using a sample of 530 languages.

[16] The sixth macro-area defined by WALS is Eurasia (29 languages), which is taken as the "default group" in our statistical analysis.

[17] See, for example, Dahl (2011).

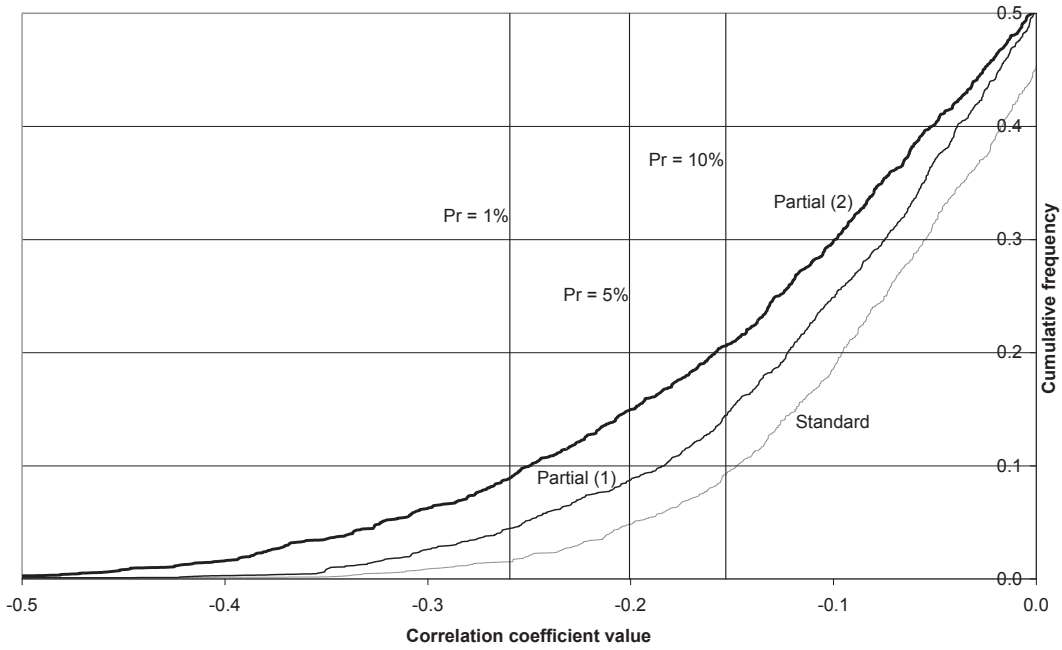


**Fig. 1.** Cumulative distribution functions for correlation coefficients.

ones (50.2%).[18] The number of significant negative coefficients is now equal to 243 (13.7%), which implies an additional increase with respect to the two previous correlation calculations.

The differences between the results of the three correlation calculations that we have performed can be appreciated by looking at the distribution functions of the correlation coefficients (Fig. 1). Those functions come from sorting the 1770 coefficients generated by each calculation from the lowest to the highest, and then evaluating the frequencies that correspond to the different values for those coefficients. The cumulative distribution functions depicted on Fig. 1 are associated with the negative coefficients generated by our three correlation calculations ("standard", "partial (1)" and "partial (2)"), and the last of them (i.e., the one that includes interactions with the 12 non-linguistic variables) displays the largest proportion of negative coefficients. That function is also the one that generates the highest frequency of significant negative coefficients for any reasonable probability level (10%, 5% or 1%).

The effects of using partial correlations and including non-linguistic variables in the calculation of the coefficients can also be illustrated by looking at a reduced set of cases like the one shown in Table 2. In that table, we have reported the standard and partial correlation coefficients for four phonological variables (the ones that correspond to features 1A, 2A, 12A and 13A), with and without the inclusion of non-linguistic variables. The outcomes of those procedures indicate that, while the only statistically significant negative standard coefficient in this set is the one that corresponds to features 12A (syllable structure) and 13A (tone), the use of partial correlation techniques (case 1) leads to a situation where the negative correlation coefficient between features 1A (consonant inventory) and 2A (vowel quality inventory) is also significant. In turn, the use of non-linguistic variables (case 2) also makes the correlation coefficient between features 1A and 13A be negative and significant. This last relationship appeared to be positive and nonsignificant when calculated as a standard correlation coefficient, and negative and nonsignificant when calculated without taking into account the possible interactions with the 12 non-linguistic variables.

Note as well that the number of negative correlation coefficients in Table 2 (both significant and nonsignificant) also increases when we move from the standard correlation case to the partial correlation cases. In the first of those situations there are only 2 negative coefficients, but this number increases to 5 when the coefficients are calculated using partial correlation techniques.

## 5. Concluding remarks

The conclusions that can be derived from the different types of analysis that we have performed using information from the 100-language WALS sample are highly dependent on which aspect of the comparisons we are most interested in. Looking

[18] These are in fact the numbers that correspond to the partial correlation coefficients between the 60 original variables. The procedure also generated 786 additional coefficients (due to the inclusion of 12 new variables), whose values cannot be compared with previous results.

**Table 2**
Correlation coefficients between selected phonological features.

| Feature | 1A | 2A | 12A | 13A |
|---|---|---|---|---|
| Standard correlation | | | | |
| 1A: Consonant inventory | 1.0000 | | | |
| 2A: Vowel quality inventory | −0.1622 | 1.0000 | | |
| 12A: Syllable structure | 0.1485 | 0.0161 | 1.0000 | |
| 13A: Tone | 0.0990 | 0.1666 | −0.3327 | 1.0000 |
| Partial correlation (case 1) | | | | |
| 1A: Consonant inventory | 1.0000 | | | |
| 2A: Vowel quality inventory | −0.3163 | 1.0000 | | |
| 12A: Syllable structure | −0.0025 | 0.2122 | 1.0000 | |
| 13A: Tone | −0.0944 | −0.0372 | −0.3180 | 1.0000 |
| Partial correlation (case 2) | | | | |
| 1A: Consonant inventory | 1.0000 | | | |
| 2A: Vowel quality inventory | −0.4108 | 1.0000 | | |
| 12A: Syllable structure | −0.0019 | 0.1693 | 1.0000 | |
| 13A: Tone | −0.3424 | −0.1763 | −0.2653 | 1.0000 |

at the standard correlation coefficients for the 60 binary variables that we have created, a possible conclusion is that complexity trade-offs is not very important, because only a few pairs of variables (4.8% of the total) display negative and statistically significant coefficients. The relative importance of complexity trade-offs, however, increases considerably when we take into account the possible interdependence among the different linguistic variables, and even more if we also control for several geographic, genetic and population-size factors. Using partial correlation techniques that allow us to include all those elements, we find that some important negative correlations clearly emerge (such as those between consonants and vowels, syllable structure and tone, alignment of verbal person marking and nominal/locational predication, number of genders and locus of marking in the clause, inflectional optative and passive voice, distance contrasts in demonstratives and noun phrase conjunction, etc.).

The existence of cross-linguistic complexity trade-offs, however, does not mean that all languages have to be equally complex in any specific sense.[19] In the 100-language WALS sample, we clearly see that some languages are complex in a large number of features, while others are simple in most of them. However, we have not found a single case in which one language dominates another one in all 60 selected features, and this means that any of the included 100 languages is more complex than any of the remaining 99 languages in at least one feature.

Nevertheless, if we want to find a case, outside the 100-language WALS sample, for which one language or dialect is completely dominated in complexity by another one, the task should not be very difficult. For example, it is very likely that a creole language is simpler than its parent language in some linguistic features and not more complex in the remaining analyzed features. What is probably impossible, however, is to find a language (creole or non-creole) that is so simple that does not surpass any other language in complexity in any linguistic feature. The only possible case like that would be a language that is simple in all the 60 features that we are considering, because for each of those features we have several examples of simple languages. That language certainly does not exist in the 100-language WALS sample (where the simplest examples have at least 14 complex features) and it probably does not exist outside the sample, either.[20]

If we look at the problem in that way, therefore, complexity trade-offs must necessarily exist, at least in some moderate degree. However, they are usually hard to find due to the interference of other (linguistic and non-linguistic) factors. Their most evident signals, however, are given by the fact that negative correlations between complexity variables increase when we control for the effect of other variables. Besides, we have also seen that, for a sufficiently long list of complexity features, there is no language that dominates any other language in all those features (especially when we compare languages that are not extremely similar, as is the case of the ones included in the WALS sample).

## Acknowledgments

I thank Damián Blasi, Guiomar Ciapuscio, Carol Fowler, Verónica Nercesián, Angelita Martínez, Elisabeth Mauder, Frans Plank and one anonymous reviewer, for their useful comments. I also thank participants at seminars held at the University of Buenos Aires and the National University of La Plata (Argentina).

[19] For a good argument about this point, see Fenk-Oczlon and Fenk (2011).
[20] The opposite extreme (i.e., a language that is complex in all 60 features) is probably equally impossible.

## Appendix 1. List of languages in the WALS sample.

| Code | Language | Macro-Area | Family |
|---|---|---|---|
| 1 | Abkhaz | Eurasia | Northwest Caucasian |
| 2 | Acoma | North America | Keresan |
| 3 | Alamblak | Papunesia | Sepik |
| 4 | Amele | Papunesia | Trans-New Guinea |
| 5 | Apurina | South America | 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 | South America | Tucanoan |
| 11 | Basque | Eurasia | Vasconic |
| 12 | Berber (Middle Atlas) | Africa | Afro-Asiatic |
| 13 | Burmese | Eurasia | Sino-Tibetan |
| 14 | Burushaski | Eurasia | Burushaskian |
| 15 | Canela-Kraho | South America | 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 | South America | Cariban |
| 36 | Hmong Njua | Eurasia | Hmong-Mien |
| 37 | Imonda | Papunesia | Border |
| 38 | Indonesian | Papunesia | Austronesian |
| 39 | Jakaltek | North America | 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 | Eurasia | 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 | Eurasia | Sino-Tibetan |
| 68 | Mixtec (Chalcatongo) | North America | Oto-Manguean |

(*continued*)

| Code | Language | Macro-Area | Family |
|---|---|---|---|
| 69 | Ngiyambaa | Australia | Pama-Nyungan |
| 70 | Oneida | North America | Iroquoian |
| 71 | Oromo (Harar) | Africa | Afro-Asiatic |
| 72 | Otomi (Mezquital) | North America | Oto-Manguean |
| 73 | Paiwan | Papunesia | Austronesian |
| 74 | Persian | Eurasia | Indo-European |
| 75 | Piraha | South America | Mura |
| 76 | Quechua (Imbabura) | South America | Quechuan |
| 77 | Rama | North America | Chibchan |
| 78 | Rapanui | Papunesia | Austronesian |
| 79 | Russian | Eurasia | Indo-European |
| 80 | Sango | Africa | Niger-Congo |
| 81 | Sanuma | South America | 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 | Eurasia | Tai-Kadai |
| 88 | Tiwi | Australia | Tiwian |
| 89 | Tukang Besi | Papunesia | Austronesian |
| 90 | Turkish | Eurasia | Altaic |
| 91 | Vietnamese | Eurasia | Austro-Asiatic |
| 92 | Warao | South America | Waraoan |
| 93 | Wari | South America | Chapacuran |
| 94 | Wichita | North America | Caddoan |
| 95 | Wichi | South America | Matacoan |
| 96 | Yagua | South America | Peba-Yaguan |
| 97 | Yaqui | North America | Uto-Aztecan |
| 98 | Yoruba | Africa | Niger-Congo |
| 99 | Zoque (Copainala) | North America | Mixe-Zoque |
| 100 | Zulu | Africa | Niger-Congo |

## Appendix 2. List of binary complexity variables.

| Code | WALS feature | Complex (variable = 1) if: |
|---|---|---|
| 1A | Consonant inventories | Consonant phoneme inventory > 25. |
| 2A | Vowel quality inventories | Vowel quality inventory > 6. |
| 4A | Voicing in plosives and fricatives | Voicing is distinctive for at least one plosive or fricative phoneme. |
| 6A | Uvular consonants | There is at least one uvular consonant phoneme. |
| 7A | Glottalized consonants | There is at least one glottalized consonant phoneme. |
| 10A | Vowel nasalization | Nasalization is distinctive for at least one vowel phoneme. |
| 11A | Front rounded vowels | There is at least one front rounded vowel phoneme. |
| 12A | Syllable structure | Syllable structure is complex. |
| 13A | Tone | Tone is distinctive. |
| 14A | Fixed stress locations | Stress location is not fixed. |
| 19A | Presence of uncommon consonants | There is at least one “uncommon” consonant phoneme (labial–velar, click, pharyngeal or /θ/). |
| 20A | Fusion of selected inflectional formatives | The language is not isolating. |
| 22A | Inflectional synthesis of the verb | Categories per word > 5. |
| 23A | Locus of marking in the clause | There is some marking in the clause. |
| 26A | Prefixing vs. suffixing in inflectional morphology | There is some inflectional morphology (prefixing, suffixing or both). |
| 28A | Case syncretism | There is some case marking. |
| 29A | Syncretism in verbal person/number marking | There is some person/number marking (syncretic or not). |
| 30A | Number of genders | Number of genders > 1. |
| 34A | Occurrence of nominal plurality | All nouns have plural forms, and their use is obligatory. |
| 37A | Definite articles | There are definite articles. |
| 38A | Indefinite articles | There are indefinite articles. |
| 39A | Inclusive/exclusive distinction in independent pronouns | There are different inclusive and exclusive plural personal pronouns. |
| 40A | Inclusive/exclusive distinction in verbal inflection | There are different inclusive and exclusive plural verbal inflections. |
| 41A | Distance contrasts in demonstratives | There are 3-way contrasts or more. |
| 44A | Gender distinctions in independent personal pronouns | There is some gender distinction in personal pronouns. |
| 45A | Politeness distinctions in pronouns | There is some politeness distinction when using second person pronouns. |

(*continued on next page*)

(*continued*)

| Code | WALS feature | Complex (variable = 1) if: |
|---|---|---|
| 47A | Intensifiers and reflexive pronouns | Intensifiers are different from reflexive pronouns. |
| 48A | Person marking on adpositions | There is some person marking. |
| 49A | Number of cases | Number of cases > 1. |
| 51A | Position of case affixes | There are some case affixes or adpositional clitics. |
| 55A | Numeral classifiers | There are numeral classifiers. |
| 58A | Obligatory possessive inflection | There is possessive inflection. |
| 59A | Possessive classification | There are two or more classes of possessive classification. |
| 63A | Noun phrase conjunction | Markers for noun phrase conjunction and commitative phrases are different. |
| 64A | Nominal and verbal conjunction | Markers for nominal and verbal conjunction are different. |
| 65A | Perfective/imperfective aspect | There is grammatical marking of the perfective/imperfective aspect. |
| 66A | The past tense | There is grammatical marking of the past tense. |
| 67A | The future tense | There is an inflectional future tense. |
| 68A | The perfect | There is a differentiated perfect tense. |
| 69A | Position of tense-aspect affixes | There is some tense-aspect inflection. |
| 70A | The morphological imperative | There is at least one form of second person imperative. |
| 73A | The optative | There is an inflectional optative. |
| 77A | Semantic distinctions of evidentiality | There are some grammatical markers of evidentiality. |
| 79A | Suppletion according to tense and aspect | There is some suppletion according to tense, aspect or both. |
| 80A | Verbal Number and suppletion | Verbs are different if the participants are singular or plural. |
| 81A | Order of Subject, object and verb | There is no dominant order. |
| 84A | Order of object, oblique, and verb | There is no dominant order. |
| 92A | Position of polar question particles | There are polar question particles. |
| 98A | Alignment of case marking of full noun phrases | The alignment is not neutral. |
| 100A | Alignment of verbal person marking | The alignment is not neutral. |
| 104A | Order of person markers on the verb | Agentive and patient arguments of the transitive verb can both occur. |
| 107A | Passive constructions | There are passive constructions. |
| 108A | Antipassive constructions | There are antipassive constructions (implicit patient or oblique patient). |
| 109A | Applicative constructions | There are applicative constructions (benefactive or non-benefactive). |
| 113A | Symmetric and asymmetric standard negation | There are different structures for the affirmative and the negative forms. |
| 119A | Nominal and locational predication | Nominal and locational predications are different. |
| 129A | Hand and arm | There are different words for “hand” and “arm”. |
| 130A | Finger and hand | There are different words for “finger” and “hand”. |
| 143E | Preverbal negative morphemes | There are preverbal negative morphemes. |
| 143F | Postverbal negative morphemes | There are postverbal negative morphemes. |

## References


Coloma, Germán, 2016. The existence of negative correlation between linguistic measures across languages. Corpus Linguist. Linguist. Theory http://dx.doi.org/10.1515/cllt-2015-0020.

Dahl, Osten, 2011. Are small languages more or less complex than big ones? Linguist. Typol. 15, 171–175.

Dryer, Matthew, 2009. Problems testing typological correlations with the online WALS. Linguist. Typol. 13, 121–135.

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

Fenk-Oczlon, Gertraud, Fenk, August, 1999. Cognition, quantitative linguistics and systemic typology. Linguist. Typol. 3, 151–177.

Fenk-Oczlon, Gertraud, Fenk, August, 2008. Complexity trade-offs between the subsystems of language. In: Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact and Change. John Benjamins, Amsterdam, pp. 43–65.

Fenk-Oczlon, Gertraud, Fenk, August, 2011. Complexity trade-offs in language do not imply an equal overall complexity. In: Solovyev, V., Polyakov, V. (Eds.), Text Processing and Cognitive Techonlogies. Kazan State University Press, Kazan, pp. 145–148.

Gingrich, Paul, 2004. Introductory Statistics for the Social Sciences. University of Regina, Regina.

Gordon, Matthew, 2016. Phonological Typology. Oxford University Press, New York.

Joseph, John, Newmeyer, Frederick, 2012. ‘All Languages are Equally Complex’: the rise and fall of a consensus. Historiogr. Linguist. 39, 341–368.

Maddieson, Ian, 2007. Issues of phonological complexity: statistical analysis of the relationship between syllable structures, segment inventories and tone contrasts. In: Solé, M., Beddor, P., Ohala, M. (Eds.), Experimental Approaches to Phonology. Oxford University Press, New York, pp. 93–103.

Matasovic, Ranko, 2014. Verbal and adnominal agreement: areal distribution and typological correlations. Linguist. Typol. 18, 171–214.

McWhorter, John, 2001. The World’s simplest grammars are Creole grammars. Linguist. Typol. 5, 125–166.

Miestamo, Matti, 2008. Grammatical complexity in a cross-linguistic perspective. In: Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact and Change. John Benjamins, Amsterdam, pp. 23–41.

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

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

Nichols, Johanna, 2009. Linguistic complexity: a comprehensive definition and survey. In: Sampson, G., Gil, D., Trudgill, P. (Eds.), Language Complexity as an Evolving Variable. Oxford University Press, Oxford, pp. 110–125.

Parkvall, Mikael, 2008. The simplicity of creoles in a cross-linguistic perspective. In: Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact and Change. John Benjamins, Amsterdam, pp. 265–285.

Prokhorov, A.V., 2002. Partial correlation coefficient. In: Hazewinkel, M. (Ed.), Encyclopedia of Mathematics. Springer, New York.

Sampson, Geoffrey, 2009. A linguistic axiom challenged. In: Sampson, G., Gil, D., Trudgill, P. (Eds.), Language Complexity as an Evolving Variable. Oxford University Press, Oxford, pp. 1–18.

Shosted, Ryan, 2006. Correlating complexity: a typological approach. Linguist. Typol. 10, 1–40.

Sinnemäki, Kaius, 2008. Complexity trade-offs in core argument marking. In: Miestamo, M., Sinnemäki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact and Change. John Benjamins, Amsterdam, pp. 67–88.

Sinnemäki, Kaius, 2014. Complexity trade-offs: a case study. In: Newmayer, F., Preston, L. (Eds.), Measuring Grammatical Complexity. Oxford University Press, New York, pp. 179–201.

Wichmann, Soren, Rama, Taraka, Holman, Eric, 2011. Phonological diversity, word length and population sizes across languages: the ASJP evidence. Linguist. Typol. 15, 157–177.