DLINE JOURNALS

International Journal of Computational Linguistics Research

Print ISSN: 0976-416X
Online ISSN: 0976-4178

IJCLR 2024; 15 (1)
https://doi.org/10.6025/jcl/2024/15/1/18-33

Language Complexity Trade-Offs: New Empirical Evidence

Germán Coloma
CEMA University
Argentina
{gcoloma@cema.edu.ar}

ABSTRACT

In this paper we provide empirical evidence to evaluate the results that originally appeared in Coloma (2015) and Coloma (2017), using a newly-assembled database of 50 languages for which we have the same text (which is the fable known as "The North Wind and the Sun"). Most conclusions of the original papers remain the same, especially the ones that signal the existence of language complexity trade-offs. This is particularly clear when we look at partial correlation coefficients between three linguistic ratios (phonemes per syllable, syllables per word, and words per clause), when we use simultaneous-equation regression methods, and when we estimate different versions of the Menzerath law, that relate phonemes per word and words per clause.

Received: 1 October 2023
Revised: 4 December 2023
Accepted: 12 December 2023

Keywords: *Complexity Trade-Off, Partial Correlation, Linguistic Ratios, Menzerath Law*

1. Introduction

In Coloma (2015) and Coloma (2017), there is an analysis focused on the possible existence of language complexity trade-offs using empirical measures. All those measures come from the text of a relatively famous fable ("The North Wind and the Sun") translated into 50 different languages. With those translations, a database was built, using information about different complexity measures for the text under analysis (phonemes per syllable, syllables per word, phonemes per word, words per clause), together with other variables related to the typological characteristics of the languages (e.g., size of the phoneme inventory, number of genders and cases, inflectional categories of the verbs) and some additional "non-linguistic" variables (e.g., location of the languages, phylogenetic characteristics, number of speakers).

The main conclusion of the abovementioned papers is that language complexity trade-offs exist and are significant in the context under analysis. They also seem to be partially hidden, because of possible interactions among different variables. As a consequence of that, it holds that the correlation and regression coefficients that relate the different variables seem to be higher and more significant when those interactions are taken into account. In order to do that, different alternative strategies were combined.They implied using partial correlation coefficients, simultaneous-regression equations, non-linguistic variables and instrumental variables.

One limitation of the analyses that appear in Coloma (2015) and Coloma (2017), however, has to do with the database itself, which consists of 50 languages (and therefore has only 50 observations). That limitation was due to the fact that, when those analyses were performed, there were relatively few sources that could be used to compare those languages, and many of those sources were about languages that were not different enough (in terms of their phylogenetic and/or geographic variation).

As several years have passed, we have been able to build another alternative database with 50 additional languages for which we have the text of "The North Wind and the Sun". The source of those languages is essentially the same one that was used for the original sample, i.e., it is the collection of "Illustrations of the IPA" published in IPA (1999) and in the Journal of the International Phonetic Association, which is now considerably larger.[1] This new database is similar to the original one, in the sense that it has languages from a variety of families, and with the same geographic distribution (ten languages from each of the five regions in which we divided the world).

In this paper, we use our newly-assembled database to perform essentially the same analyses that appear in Coloma (2015) and Coloma (2017). First, we describe the basic characteristics of the database in terms of its scope of languages and the value of the calculated complexity measures (section 2). Then, in section 3, we use those measures to calculate correlation coefficients, using alternative methodologies. In section 4, we estimate different versions of the so-called "Menzerath law", which proposes a negative relationship between phonemes per word and words per clause. Later on, in section 5, we compare the new results with the ones that appear in Coloma (2015) and Coloma (2017). Finally, in section 6, we state a few concluding remarks.

2. The North Wind and the Sun

The fable of the North Wind and the Sun, attributed to Aesop, is a text that has been used for many decades by the International Phonetic Association as a "specimen" or model to illustrate the sounds of languages, and also the phonetic symbols that are suitable to describe those sounds.[2] It is therefore a unique case of a short text for which specialists in the phonetics of different languages have analyzed the sounds, the phonemes, the syllables and the words of the languages and dialects under study.

In Coloma (2015) and Coloma (2017), the observations come from a database that relies on the text of "The North Wind and the Sun" translated into the following languages: Sahaptin, Apache, Chickasaw, Seri, Trique, Zapotec, Ecuadorian Quichua, Shiwilu, Yine and Mapudungun (which are original of the American continent); Portuguese, Spanish, Basque, French, Irish, English, German, Russian, Hungarian and Greek (from Europe); Tashlhiyt Berber, Temne, Kabiye, Igbo, Hausa, Dinka, Nara, Amharic, Sandawe and Bemba (from Africa); Georgian, Turkish, Hebrew, Standard Arabic, Persian, Tajik, Nepali, Hindi, Bengali and Tamil (from West Asia); and Japanese, Korean, Mandarin, Cantonese, Burmese, Thai, Vietnamese, Malay, Tausug and Arrernte (from East Asia and Australasia).

For this paper, we have built a new database with 50 additional languages (see figure 1). The languages included are: Gitksan, Paiute, Kumiai, Amuzgo, Qanjobal, Aingae, Urarina, Shawi, Shipibo and Cusco Quechua (America); Scottish Gaelic, Galician, Catalan, Italian, Croatian, Dutch, Swedish, Polish, Estonian and Ukrainian (Europe); Zwara Berber, Seenku, Ibibio, Tera, Kera, Kunama, Shilluk, Lusoga, Malagasy and Setswana (Africa); Kazakh, Azerbaijani, Khuzestani Arabic, Kumzari, Dari, Punjabi, Sumi, Assamese, Telugu and Malayalam (West Asia); and Shanghainese, Xiang, Hmong, Lizu, Sama, Mah Meri, Madurese, Nen, Pitjantjatjara and Hawaiian (East Asia and Australasia).

---

[1]We also included three examples from additional sources. Two of them (Amuzgo and Cusco Quechua) are taken from a collection of phonetic illustrations published by Marlett (2009), and a third one (Qanjobal) appeared as a working paper of the University of Illinois (Lichtman et al, 2010).

[2]See, for example, IPA (1949) and IPA (1999).

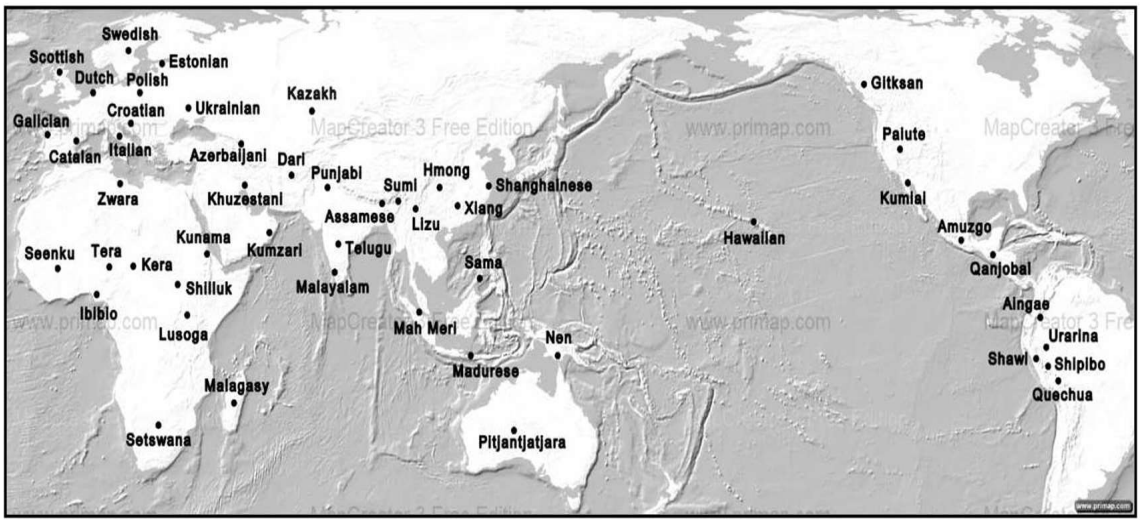

Figure 1. Location of the languages included in the sample

The basic statistics computed for this sample of languages come from counting the number of phonemes, syllables, words and clauses included in the translation of "The North Wind and the Sun" for each of those languages. With those figures, we calculated a series of linguistic ratios, which basically are the phoneme/syllable ratio, the syllable/word ratio and the word/clause ratio. The phoneme/syllable ratio goes from a minimum of 1.7905 (for the case of Shipibo, a Panoan language spoken in Peru) to a maximum of 2.9024 (for the case of Kumiai, a Yuman language spoken in the Mexican/US border), in a context in which the average number of phonemes per syllable is 2.2957. The syllable/word ratio, conversely, goes from a minimum of 1.0637 (for Hmong, a Hmong-Mien language spoken in China) to a maximum of 3.6 (for Telugu, a Dravidian language spoken in India), in a context where the average number of syllables per word is 2.1202. The minimum word/clause ratio, in turn, is equal to 4.5, and it corresponds to Paiute (which is a Uto-Aztecan language spoken in the US), while the maximum word/clause ratio in the sample is 23.83, and it corresponds to the Hawaiian language (in a context in which the average number of words per clause is 11.25).[3]

3. Standard and Partial Correlation Coefficients

The easiest way to detect possible trade-offs between language complexity measures is to calculate correlation coefficients between the linguistic ratios mentioned in the previous section of this paper. As in this case we have three main ratios (phonemes per syllable, syllables per word, and words per clause), it is possible to find three basic measures of correlation, which are the ones that appear on table 1.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1,0000		
Syllables per word	-0,4202	1,0000	
Words per clause	-0,2299	-0,4697	1,0000

Table 1. Standard correlation coefficients.

[3]*For the complete list of the values of the linguistic ratios, see Appendix 1.*

The basic meaning of the reported correlation coefficients is that the value of each of the calculated variables (which can be seen as empirical measures of partial language complexity) is negatively correlated with the other two variables. This gives a hint of possible trade-offs, in the sense that it implies that, on average, a language that is more complex in a certain dimension tends to be simpler in another dimension. For example, in this database it holds that the text of "The North Wind and the Sun" translated into Scottish Gaelic has an average of 1.30 syllables per word and an average of 23.33 words per clause. Conversely, the same text in Cusco Quechua has an average of 3.026 syllables per word, but only 8.56 words per clause. This could be seen as an illustration that languages that tend to use longer (and more complex) words generally use fewer words per clause (and they probably have simpler sentences).

The absolute values of the correlation coefficients are also related to the statistical significance of those coefficients. For example, correlation between syllables per word and words per clause (r = -0.4697) is significantly different from zero at a 1% probability level (p = 0.0006), and the same occurs with correlation between phonemes per syllable and syllables per word (p = 0.0024). Conversely, correlation between phonemes per syllable and words per clause, though negative, fails to be significant at any reasonable probability level, since its corresponding "p-value" (p = 0.1082) is above 10%.

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

A partial correlation coefficient, therefore, is a measure of the linear dependence for a pair of variables in the case where the influence of other variables is suppressed. To calculate that coefficient, it is necessary to control for the possible effect of other factors on the two variables that we wish to correlate, and to eliminate that effect using some statistical procedure. One possibility is to begin with a complete correlation matrix for all the variables under analysis (which in our case are only three variables), and then invert that matrix. Once we do that, we can use the following formula:

$$r = -\frac{P_{xy}}{\sqrt{P_{xx} P_{yy}}} \qquad (1)$$

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.[4] This process of matrix inversion is actually one of the possibilities that can be used to obtain partial correlation coefficients. Another one is to run a set of three regression equations, in which each variable is regressed against a constant and the other two variables. Both procedures have the same goal, which is pulling out the effects that the remaining variable may have on each pair of variables that we are interested in.

If we apply the regression procedure in this case, we need to run a system of regression equations that consists of the following functions:

$$Phon/Syll = c(1) + c(2)*Syll/Word + c(3)*Word/Clause \qquad (2)$$

$$Syll/Word = c(4) + c(5)*Phon/Syll + c(6)*Word/Clause \qquad (3)$$

$$Word/Clause = c(7) + c(8)*Phon/Syll + c(9)*Syll/Word \qquad (4)$$

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

where *Phon/Syll*,*Syll/Word* and *Word/Clause* are our three linguistic ratios, and *c(1)*, *c(2)*, *c(3)*, *c(4)*, *c(5)*, *c(6)*, *c(7)*, *c(8)* and *c(9)* are the coefficients to be estimated.

Concept	Coefficient	t-Statistic	Probability
Phoneme/Syllable equation			
Constant [c(1)]	3,352490	17,764440	0,0000
Syllable/Word [c(2)]	-0,288498	-5,343526	0,0000
Word/Clause [c(3)]	-0,039564	-4,323088	0,0000
R-squared	0,4108		
Syllable/Word equation			
Constant [c(4)]	6,267824	9,768979	0,0000
Phoneme/Syllable [c(5)]	-1,309963	-5,343526	0,0000
Word/Clause [c(6)]	-0,101361	-5,729814	0,0000
R-squared	0,5152		
Word/Clause equation			
Constant [c(7)]	36,361480	7,786741	0,0000
Phoneme/Syllable [c(8)]	-7,191039	-4,323088	0,0000
Syllable/Word [c(9)]	-4,057344	-5,729814	0,0000
R-squared	0,4424		

Table 2. Regression results to calculate partial correlation coefficients

When we run that system of regression equations using ordinary least squares,[5] we obtain the results that appear on table 2. With those regression coefficients, the partial correlations between the different linguistic ratios can be calculated by using the following formula:

$$r_{12} = -\sqrt{\beta_{12}.\beta_{21}} \qquad (5)$$

where r_{12} is the partial correlation coefficient between variable 1 and variable 2, β_{12} is the regression coefficient of variable 2 in variable 1's equation, and β_{21} is the regression coefficient of variable 1 in variable 2's equation. Note that in this formula we assume that, as both regression coefficients are negative, the corresponding partial correlation coefficient must also be negative.

Applying our formula to the results reported on table 2, it is possible to obtain the partial correlation coefficients that are shown on table 3. If we compare those results with the ones that appear on table 1, we see that the three partial correlation coefficients are higher than their corresponding standard correlation coefficients. This is also linked to a larger statistical significance, which in this case is given by the fact that the three calculated coefficients are now significant at a 1% probability level (p = 0.0000; p = 0.0001 and p = 0.0000).

[5]These regressions, and all the others whose results appear in this paper, were run using the statistical program EViews 10.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
Phonemes per syllable	1,0000		
Syllables per word	-0,6148	1,0000	
Words per clause	-0,5334	-0,6413	1,0000

Table 3. Partial correlation coefficients

In Coloma (2017), ther is also an exploration of the possibility to calculate partial correlation coefficients that control for other additional variables, related to geographic, phylogenetic and population factors. This can be done by running a regression-equation system that includes those additional variables, such as the following one:

$$\begin{aligned} Phon/Syll = {} & c(1)*Europe + c(2)*Africa + c(3)*Westasia + c(4)*Eastasia \\ & + c(5)*America + c(6)*Indoeuro + c(7)*Afroasiatic + c(8)*Nigercongo \\ & + c(9)*Sinotibetan + c(10)*Austronesian + c(11)*Major \\ & + c(12)*Syll/Word + c(13)*Word/Clause \end{aligned} \quad (6)$$

$$\begin{aligned} Syll/Word = {} & c(21)*Europe + c(22)*Africa + c(23)*Westasia + c(24)*Eastasia \\ & + c(25)*America + c(26)*Indoeuro + c(27)*Afroasiatic + c(28)*Nigercongo \\ & + c(29)*Sinotibetan + c(30)*Austronesian + c(31)*Major \\ & + c(32)*Phon/Syll + c(33)*Word/Clause \end{aligned} \quad (7)$$

$$\begin{aligned} Word/Clause = {} & c(41)*Europe + c(42)*Africa + c(43)*Westasia + c(44)*Eastasia \\ & + c(45)*America + c(46)*Indoeuro + c(47)*Afroasiatic + c(48)*Nigercongo \\ & + c(49)*Sinotibetan + c(50)*Austronesian + c(51)*Major \\ & + c(52)*Phon/Syll + c(53)*Syll/Word \end{aligned} \quad (8)$$

where Europe, Africa, Westasia, Eastasia and America are binary variables that take a value equal to one when a language belongs to a certain area (and zero otherwise); Indoeuro, Afroasiatic, Nigercongo, Sinotibetan and Austronesian are binary variables that take a value equal to one when a language belongs to a certain linguistic family; and Major is a binary variable that takes a value equal to one when a language is spoken by more than 5 million people.[6]

Due to the fact that the determinants of equations 6, 7 and 8 are basically the same, this is a case in which our analysis can be improved if we use "simultaneous-equation regressions". This method is relatively widespread in some social sciences such as economics, since it allows for procedures that single-equation regression analyses cannot deal with. The main one is the use of the correlations between the residuals of the three regression equations, through the so-called "seemingly unrelated regression" (SUR) procedure. It implies that, when estimating one equation, we also use information from the other equations, and that information can improve the precision and the statistical efficiency of the estimated coefficients.[7]

Equations 6, 7 and 8 can be run simultaneously, to see if we can find any statistical significance

[6]*Due to this definition, the " major languages" in our sample are Assamese, Azerbaijani, Catalan, Croatian, Cusco Quechua, Dari, Dutch, Estonian, Ibibio, Italian, Kazakh, Madurese, Malagasy, Malayalam, Polish, Punjabi, Setswana, Shanghainese, Swedish, Telugu, Ukrainian and Xiang.*

[7]*This procedure was originally proposed by Zellner (1962). It is used in Coloma (2014) and Coloma (2017).*

for the coefficients labeled as c(12), c(13), c(32), c(33), c(52) and c(53), which are the ones that measure the relationships between the different linguistic ratios. That analysis was performed using both ordinary least squares (OLS)and SUR. Applying the same procedure described for equations 2, 3 and 4, we used its results to calculate new partial correlation coefficients, which are the ones reported on table 4.

Variable	Phoneme/Syllable	Syllable/Word	Word/Clause
OLS Regression			
Phonemes per syllable	1,0000	1,0000	
Syllables per word	-0,6036	-0,5340	
Words per clause	-0,4354		1,0000
SUR Regression			
Phonemes per syllable	1,0000	1,0000	
Syllables per word	-0,9009	-0,8526	
Words per clause	-0,7789		1,0000

Table 4. Partial correlation coefficients from simultaneous-equation regressions

Note that the coefficients obtained when we use SUR are in all cases higher than the ones that we find when we use OLS (and they are also larger than the coefficients reported on tables 1 and 3). This may be seen as a signal that the true negative correlation between the different linguistic ratios is higher than the one obtained when we perform a less sophisticated analysis.

4. The Menzerath Law

The Menzerath law states that the length of a linguistic construct is an inverse function of the length of the construct's constituents. Originally proposed by Menzerath (1954), this law was reformulated by Altmann (1980) as a power function that can be written in the following way:

$$y = a \cdot x^b \tag{9}$$

where y is the average length of a linguistic construct, measured in its constituents, x is the average length of the construct's constituents, measured in their subconstituents, a is a positive parameter, and b is a negative parameter.[8]

In a more recent paper (Milicka, 2014), it is argued that the power function formula for the Menzerath law can be replaced by a hyperbolic function, written in the following way:

$$y = a + \frac{b}{x} \tag{10}$$

where a and b are both positive. This formula is supposed to fit some datasets better and to have a more intuitive explanation, related to a trade-off between plain information and structure information (Köhler, 1984). In Coloma (2015), the same 50-language sample of Coloma (2017)

[8] *In fact, Altmann's formula also includes an additional exponential term ($e^{c.x}$). This term disappears when we solve the formula as a differential equation.*

was used to explore the implications of Menzerathlaw . In particular, the regression equations 9 and 10 were run, using words per clause as a measure of variable y and phonemes per word as a measure of variable x. The idea is that clauses are linguistic constructs whose main constituents are words, while words are constituents whose main subconstituents are phonemes.

The basic conclusion obtained in Coloma (2015) is that both the power function and the hyperbolic function perform well to explain the strong negative correlation that exists between phonemes per word and words per clause in the context under analysis, and that there is no evidence to assess that the hyperbolic alternative is actually better than the original power function formulation proposed by Altmann (1980). The same analysis can be performed using our newly-assembled database, for which we can run regression equations such as the following:

$$Ln\ (Word/Clause) = c(1) + c(2)*Ln\ (Phon/Word) \quad (11)$$

$$Word/Clause = c(3) + c(4)*[1/(Phon/Word)] \quad (12)$$

These formulae are linear versions of equations 9 and 10, for the case where the independent variable is a logarithmic or an inverse transformation of the phoneme/word ratio (Phon/Word), and the dependent variable is the word/clause ratio (Word/Clause) or its logarithmic transformation.

The main results for those regressions, run using ordinary least squares, appear on table 5. In it, we can see that both specifications generate a relatively good fit for the data, and the estimated regression coefficients are also highly significant and have the expected signs (since they both imply a negative relationship between Word/Clause and Phon/Word). Based on the R^2 coefficients, we can also find that the fit of the power function (R^2 = 0.3747) is slightly worse than the one obtained with the hyperbolic function (R^2 = 0.4111).[9]

Concept	Coefficient	t-Statistic	Probability
Power function			
Constant [c(1)]	3,475605	16,677250	0,0000
Phon/Word [c(2)]	-0,717514	-5,363109	0,0000
R-squared	0,3747		
Hyperbolic function			
Constant [c(3)]	2,858995	1,902256	0,0631
Phon/Word [c(4)]	37,647220	5,788830	0,0000
R-squared	0,4111		

Table 5. Regression results for Menzerath law's OLS estimations

The power-function and hyperbolic-function regression equations can also be graphed in a diagram in which we represent the different language observations in terms of phonemes per word versus words per clause. This is what appears on figure 2, in which we see that the hyperbolic regression equation predicts a value for Word/Clause that is always higher than the one predicted by the power-function equation. This generates a better fit for 25 languages (e.g., Hawaiian, Catalan, Punjabi, Italian, Telugu) but a worse fit for the remaining 25 languages (e.g., Hmong, Croatian, Polish, Paiute, Azerbaijani).

[9] *Both specifications also have a better fit than the one that could be obtained under a simpler linear specification. That specification would have produced an R^2 coefficient equal to 0.3481.*

The results reported on table 5 (and depicted on figure 2) are nevertheless subject to some possible criticism, due to the fact that they are produced by OLS regressions that depend on several statistical assumptions that do not necessarily hold in the context under analysis. This has to do with the fact that, when one performs a regression between two variables, it is implicitly assumed that the variable on the right-hand side of the equation (i.e., the independent variable) is the one that explains the behavior of the variable on the left-hand side of the equation (i.e., the dependent variable), and not the other way round. This is a noticeable difference between regression and correlation analyses, since correlation is a symmetrical concept that assumes no particular causal direction from one variable to the other.

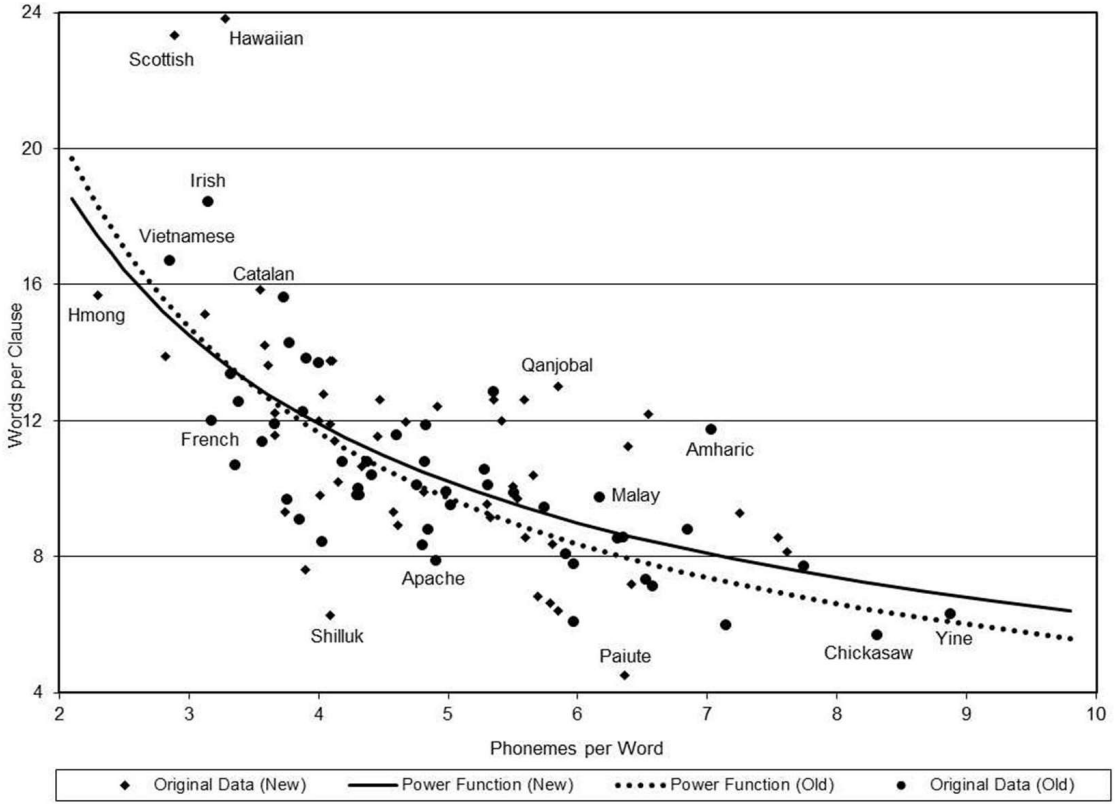

Figure 2. Power and hyperbolic regression lines

In the case under study in this paper, the logic of the Menzerath law indicates that the nature of the constituents of a language (i.e., the number of phonemes per word) determines the structure of the higher-level construct (i.e., the number of words per clause). However, this causality is not completely clear in our problem, since we are examining a cross-linguistic context where the relationship between the two variables can be interpreted as a signal of the existence of a complexity trade-off. In that context, both the word/clause ratio and the phoneme/word ratio may be variables that are simultaneously determined by an external process.

To deal with this kind of endogeneity issues we can use instrumental variables, i.e., variables that are supposed to be related with the independent variable under analysis but have the property that they are determined exogenously (i.e., outside the statistical problem that we are analyzing). For this particular case, we have chosen to use the eleven binary variables introduced in the previous section to deal with geographic, phylogenetic and population factors (Europe, Africa, Westasia, Eastasia, America, Indoeuro, Afroasiatic, Nigercongo, Sinotibetan, Austronesian and Major), together with six "typological variables" that come from the different languages' grammars. Those variables are the number of consonant phonemes in each language's inventory (Consonants), the number of vowel phonemes in that inventory (Vowels), the number of distinctive tones that each language possesses (Tones), the number of distinctive genders that nouns may

have (Genders), the number of distinctive cases for those nouns (Cases), and the number of inflectional categories of the verbs (Inflections).[10] The figures for the first three typological variables are taken from the same sources used to obtain the different versions of "The North Wind and the Sun" (i.e., from the corresponding illustrations of the IPA). To impute values for the last three variables, conversely, we used the online version of the World Atlas of Language Structures (WALS), edited by Dryer & Haspelmath (2013).

In a case like this, one can use a procedure to include the instrumental variables in the estimation of the equation coefficients that is known as "two-stage least squares" (2SLS). It consists of a first stage in which the endogenous independent variable (in our case, Phon/Word) is regressed against all the instrumental variables, using ordinary least squares. Then there is a second stage in which the fitted values of that regression are included in the estimation of the actual equation that one wishes to regress (in our case, in each of the Menzerath law equations), instead of the original values for the endogenous independent variable.[11]

Concept	Coefficient	t-Statistic	Probability
Power function			
Constant [c(1)]	3,430491	13,262640	0,0000
Phon/Word [c(2)]	-0,688139	-4,123453	0,0001
R-squared	0,3741		
Hyperbolic function			
Constant [c(3)]	3,805675	1,965397	0,0552
Phon/Word [c(4)]	33,400260	3,929041	0,0003
R-squared	0,4059		

Table 6. Regression results for Menzerath law's 2SLS estimations

The results of these regressions are reported on table 6. That table shows that the corresponding R^2 coefficients are slightly smaller than the ones reported on table 5. This has to do with the fact that an estimation that uses instrumental variables is always less efficient than another estimation that uses the original variables, although it can be more consistent (i.e., closer to the true values of the parameters that would be obtained if one knew the whole set of data that is generating the process under estimation).

For the case of the 2SLS coefficients shown on table 6, the results are in line with the estimations performed using OLS, in the sense that the estimated parameters are significantly different from zero and imply a negative relationship between phonemes per word and words per clause. Once again, the hyperbolic function has a small advantage in terms of goodness of fit over the power function, since "R^2(Hyperbolic) = 0.4059" while "R^2(Power) = 0.3741".

5. Comparison with previous results

The results reported in the two previous sections, obtained using a newly-assembled database of

[10] *To see the values of these variables in each of the languages, see Appendix 2.*

[11] *This procedure was originally proposed by Basmann (1957). For a more complete explanation, see Davidson & MacKinnon (2003), chapter 8.*

50 languages, can be compared with the original results that appear in Coloma (2015) and Coloma (2017). Performing that comparison (see table 7), we can see that several stylized facts remain the same. For example, for both samples it holds that the partial correlation coefficients are higher than their corresponding standard correlation coefficients, and that those coefficients increase even more when we use an estimation method based on seemingly unrelated regressions (SUR). It also occurs that the "ranking" of the correlation coefficients is unaltered (since the highest coefficient is the one that relates Syll/Word with Word/Clause, followed by the coefficient that relates Phon/Syll with Syll/Word, while the coefficient that relates Phon/Syll with Word/Clause is always the one with the lowest absolute value).

The basic similarity between the two databases, as we already mentioned in section 2, is the fact that they both have 50 observations, and that the basic geographic division is the same (10 languages from each regions of the world, which are America, Europe, Africa, West Asia and East Asia, including Australasia). The old database is slightly more diverse geographically in America, but the new one is certainly more diverse in Australasia, since it includes one language from New Guinea (Nen) and another one from Polynesia (Hawaiian). In the old database, there are a few languages from families that are not represented in the new database (such as Apache, Mapudungun, Basque, Sandawe and Georgian), but the new database also has languages whose families do not appear in the old database (such as Paiute, Qanjobal, Shipibo, Hmong and Nen). The old database has a considerably larger proportion of "major languages" (58% versus 44%),[12] basically because it includes almost all the languages spoken by more than 100 million people (Mandarin, English, Spanish, Hindi, Arabic, Portuguese, Russian, Bengali, Japanese).

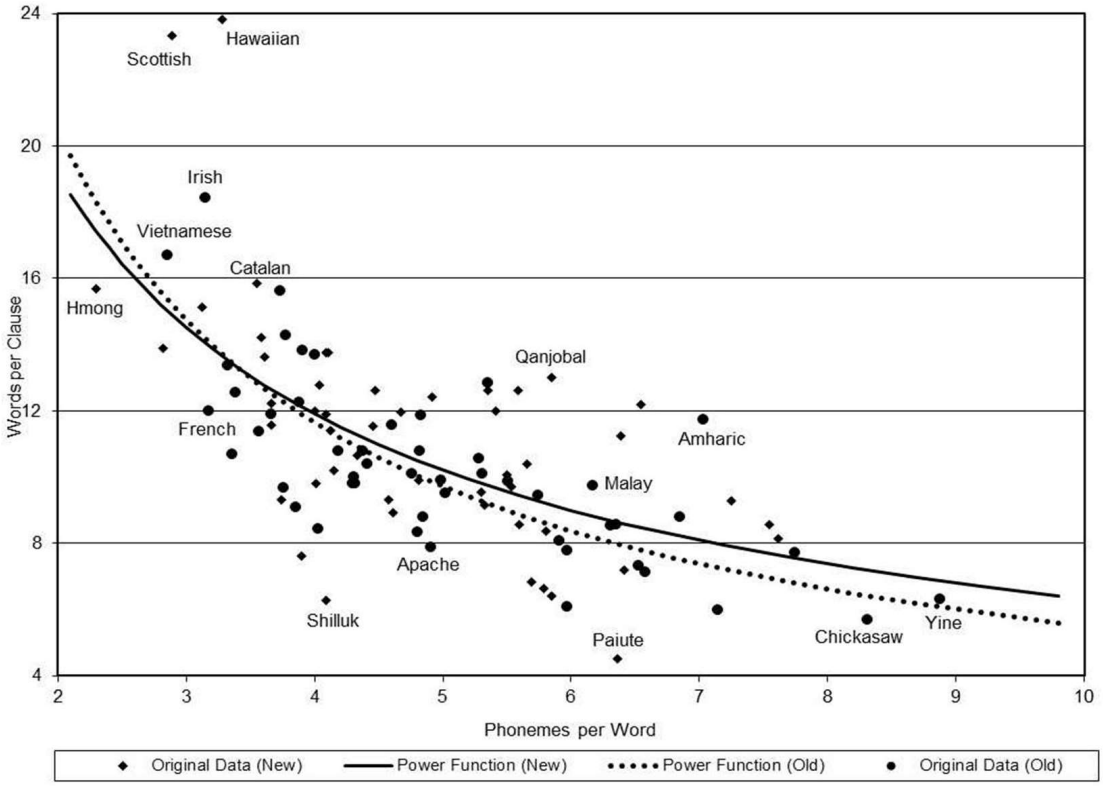

Figure 3. Menzerath law regression lines

[12]*This is not necessarily good for a sample of languages, since it is estimated that only 182 languages (2.6%) are used by more than 5 million people, from a total of 7117 languages spoken around the world (Eberhard, Simons &Fennig, 2020).*

	Old database		New database	
Concept	**Coefficient**	**Probability**	**Coefficient**	**Probability**
Correlation coefficients				
Standard correlation				
Phon/Syl vs. Syl/Word	-0,2420	0,0905	-0,4202	0,0024
Phon/Syl vs. Word/Clause	-0,0522	0,7187	-0,2299	0,1082
Syl/Word vs. Word/Clause	-0,6785	0,0000	-0,4697	0,0006
Partial correlation (1)				
Phon/Syl vs. Syl/Word	-0,3781	0,0074	-0,6148	0,0000
Phon/Syl vs. Word/Clause	-0,3036	0,0340	-0,5334	0,0001
Syl/Word vs. Word/Clause	-0,7132	0,0000	-0,6413	0,0000
Partial correlation (2)				
Phon/Syl vs. Syl/Word	-0,3320	0,0340	-0,6036	0,0000
Phon/Syl vs. Word/Clause	-0,1761	0,2708	-0,4354	0,0016
Syl/Word vs. Word/Clause	-0,6330	0,0000	-0,5340	0,0001
Partial correlation (SUR)				
Phon/Syl vs. Syl/Word	-0,5852	0,0001	-0,9009	0,0000
Phon/Syl vs. Word/Clause	-0,4163	0,0068	-0,7789	0,0000
Syl/Word vs. Word/Clause	-0,8990	0,0000	-0,8526	0,0000
Menzerath law				
OLS regressions				
Power function				
Constant	3,4528	0,0000	3,4756	0,0000
Variable	-0,7310	0,0000	-0,7175	0,0000
Hyperbolic function				
Constant	2,5735	0,0122	2,8590	0,0631
Variable	36,1152	0,0000	37,6472	0,0000
2SLS regressions				
Power function				
Constant	3,5860	0,0000	3,4305	0,0000
Variable	-0,8158	0,0000	-0,6881	0,0001
Hyperbolic function				
Constant	2,1803	0,0726	3,8057	0,0552
Variable	38,2906	0,0000	33,4003	0,0003

Table 7. Comparison of results

The old and the new databases also have relatively similar distributions of languages based on the values of their linguistic ratios (which in all cases are calculated using the text of "The North Wind

and the Sun"). A glimpse of that can be seen by looking at the graph that appears on figure 3, in which we show the corresponding power function regression lines for our version of Menzerath law (i.e., words per clause versus phonemes per word), together with the original observations from the old database (circles) and the new database (rhombs).

As we can see on figure 3, most observations for both the old and the new databases are concentrated in the area in which the texts have an average of 3 to 7 phonemes per word, and an average of 7 to 16 words per clause. Nevertheless, the old database has two outliers with more than 8 phonemes per word (that correspond to Chickasaw, a Muskogean language spoken in the US; and to Yine, an Arawakan language spoken in Peru), while the new database has two outliers with more than 20 words per clause (that correspond to Hawaiian and to Scottish Gaelic). This is probably why the new database regression line is higher in the region of the graph with fewer phonemes per word, while the old database regression line is higher in the region with a larger number of phonemes per word.

6. Concluding Remarks

After performing different kinds of calculations and estimations with our newly-assembled database of languages for which we have the text of "The North Wind and the Sun", and comparing those calculations and estimations with the ones obtained for the original database used in Coloma (2015) and Coloma (2017), it is possible to derive a series of conclusions and comments.

The main conclusion is that the language complexity trade-offs that were detected in the original studies also appear in this paper (in which we use different data). Moreover, the fact that those trade-offs are more evident when we use methods that deal with the interaction among different variables remains unaltered, as can be seen when we compare standard correlation coefficients with partial correlation coefficients (which are even more significant if we use a simultaneous-equation regression method such as SUR). Some results are also statistically similar when we compare the old and the new databases. This holds for most correlation coefficients, and also for the regression coefficients derived when we estimate different alternatives for the Menzerath law.

The main differences between the original results and the newly-obtained ones, however, are the following:

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

***b)** Two partial correlation coefficients, calculated using the SUR procedure, are significantly different when computed using the old and the new databases (the ones that relate phonemes per syllable with syllables per word, and phonemes per syllable with words per clause).*

***c)** The power-function specification of the Menzerath law has a better fit with data from the old database, but a worse fit with data from the new database (compared to the fit obtained when using the hyperbolic function).*

Acknowledgements

I thank Matthew Gordon, Stefan Gries and Mirka Rauniomaa for their comments on previous versions of this paper. I also thank Oraimar Socorro, for her help in finding many of the papers that were used as data sources for this article.

References

[1] Altmann, Gabriel. (1980). Prolegomena to Menzerath's Law. *Glottometrika, 2*, 1-10.

[2] Basmann, Robert. (1957). A Generalized Classical Method of Linear Estimation of Coefficients in a Structural Equation. *Econometrica, 25*, 77–83.

[3] Coloma, Germán. (2014). Towards a Synergetic Statistical Model of Language Phonology.

Journal of Quantitative Linguistics, 21, 100-122.

[4] Coloma, Germán. (2015). The Menzerath-Altmann law in a cross-linguistic context. *SKY Journal of Linguistics, 28*, 139-159.

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

[6] Davidson, Russell, & MacKinnon, James. (2003). Econometric Theory and Methods. New York: Oxford University Press.

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

[8] Eberhard, David, Simons, Gary, & Fennig, Charles. (2020). Ethnologue: Languages of the World, 23rd edition. Dallas, SIL International.

[9] IPA. (1949). Principles of the International Phonetic Association. London: University College.

[10] IPA. (1999). Handbook of the International Phonetic Association. Cambridge: Cambridge University Press.

[11] Köhler, Reinhard. (1984). Zur Interpretation des Menzerathschen Gesetzes. *Glottometrika, 6*, 177-183.

[12] Lichtman, Karen, Chang, Shawn, Kramer, Jennifer, Crespo, Claudia, Hallett, Jill, Huensch, Amanda, & Morales, Alexandra. (2010). IPA Illustration of Q'anjob'al. *Studies in the Linguistic Sciences, University of Illinois*.

[13] Marlett, Stephen. (2009). Ilustraciones fonéticas de lenguas amerindias. Lima: SIL International.

[14] Menzerath, Paul. (1954). Die Architektonik des deutschen Wortschatzes. Bonn: Dümmler.

[15] Milicka, Jiri. (2014). Menzerath's Law: The Whole is Greater than the Sum of its Parts. *Journal of Quantitative Linguistics, 21*, 85-99.

[16] Prokhorov, A. V. (2002). Partial Correlation Coefficient. In M. Hazewinkel (Ed.), *Encyclopedia of Mathematics*. New York: Springer.

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

Appendix 1: Linguistic ratios from the assembled database

Language	Region	Family	Phon/Syl	Syl/Word	Phon/Word	Word/Claus
Aingae	America	Cofan	2,0260	2,8657	5,81	8,38
Amuzgo	America	Oto-Manguean	2,4424	1,5000	3,66	12,22
Assamese	West Asia	Indo-European	2,1142	2,6186	5,54	9,70
Azerbijani	West Asia	Altaic	2,2593	2,5200	5,69	6,82
Catalan	Europe	Indo-European	2,3735	1,4955	3,55	15,86
Croatian	Europe	Indo-European	2,2152	2,0841	4,62	8,92
Cusco Quechua	America	Quechuan	2,4936	3,0260	7,55	8,56
Dari	West Asia	Indo-European	2,4737	1,6667	4,12	11,40
Dutch	Europe	Indo-European	2,6503	1,5093	4,00	12,00
Estonian	Europe	Uralic	2,6057	2,0349	5,30	9,56
Galician	Europe	Indo-European	2,2983	1,8854	4,33	10,67
Gitksan	America	Penutian	2,7212	1,6378	4,46	11,55
Hawaiian	East Asia	Austronesian	1,9300	1,6993	3,28	23,83
Hmong	East Asia	Hmong-Mien	2,1617	1,0637	2,30	15,70
Ibibio	Africa	Niger-Congo	2,0000	2,0545	4,11	13,75
Italian	Europe	Indo-European	2,3085	1,7478	4,03	12,78
Kazakh	West Asia	Altaic	2,5785	2,4778	6,39	11,25
Kera	Africa	Afro-Asiatic	2,3935	1,5650	3,75	9,32
Khuzestani Arabic	West Asia	Afro-Asiatic	2,4770	2,2597	5,60	8,56
Kumiai	America	Yuman	2,9024	1,3443	3,90	7,63
Kumzari	West Asia	Indo-European	2,3478	1,5617	3,67	11,57
Kunama	Africa	Nilo-Saharan	2,1337	3,0656	6,54	12,20
Lizu	East Asia	Sino-Tibetan	2,0930	1,9545	4,09	13,75
Lusoga	Africa	Niger-Congo	1,8968	2,9808	5,65	10,40
Madurese	East Asia	Austronesian	2,1445	2,6040	5,58	12,63
Mah Meri	East Asia	Austro-Asiatic	2,5385	1,6364	4,15	10,21
Malagasy	Africa	Austronesian	2,0435	2,1905	4,48	12,60
Malayalam	West Asia	Dravidian	2,1951	2,6623	5,84	6,42
Nen	East Asia	Papuan	2,3021	2,3267	5,36	12,63
Paiute	America	Uto-Aztecan	2,1604	2,9444	6,36	4,50
Pitjantjatjara	East Asia	Pama-Nyungan	2,1792	2,9444	6,42	7,20
Polish	Europe	Indo-European	2,7089	1,7753	4,81	9,89
Punjabi	West Asia	Indo-European	2,2644	1,5963	3,61	13,63
Qanjobal	America	Mayan	2,3750	2,4615	5,85	13,00
Sama	East Asia	Austronesian	2,3453	2,3435	5,50	10,08
Scottish Gaelic	Europe	Indo-European	2,2198	1,3000	2,89	23,33
Seenku	Africa	Niger-Congo	2,1078	1,3360	2,82	13,89
Setswana	Africa	Niger-Congo	1,9188	1,6281	3,12	15,13
Shanghainese	East Asia	Sino-Tibetan	2,4483	1,8710	4,58	9,30
Shawi	America	Kawapanan	2,1312	3,4000	7,25	9,29
Shilluk	Africa	Nilo-Saharan	2,3804	1,7196	4,09	6,29
Shipibo	America	Panoan	1,7905	2,6079	4,67	11,95
Sumi	West Asia	Sino-Tibetan	1,8448	2,6667	4,92	12,43
Swedish	Europe	Indo-European	2,5917	1,5794	4,09	11,89
Telugu	West Asia	Dravidian	2,1154	3,6000	7,62	8,13
Tera	Africa	Afro-Asiatic	2,2390	1,6016	3,59	14,22
Ukrainian	Europe	Indo-European	2,5000	2,1667	5,42	12,00
Urarina	America	Urarinian	1,9349	2,9912	5,79	6,65
Xiang	East Asia	Sino-Tibetan	2,5192	1,5918	4,01	9,80
Zwara Berber	Africa	Afro-Asiatic	2,8898	1,8438	5,33	9,14
Average			**2,2957**	**2,1202**	**4,80**	**11,25**

Appendix 2: Typological variables

Language	Consonants	Vowels	Tones	Cases	Genders	Inflections
Aingae	27	10	1	6	1	6
Amuzgo	22	14	3	1	1	6
Assamese	39	8	1	6	2	2
Azerbijani	70	9	1	6	1	6
Catalan	25	7	1	1	2	4
Croatian	30	10	1	5	3	4
Cusco Quechua	29	5	1	8	1	8
Dari	33	8	1	2	1	4
Dutch	18	19	1	1	3	2
Estonian	17	18	1	10	1	2
Galician	27	7	1	1	2	4
Gitksan	17	9	1	4	1	10
Hawaiian	31	10	1	1	1	6
Hmong	25	8	8	1	1	2
Ibibio	17	12	2	1	1	8
Italian	31	7	1	1	2	4
Kazakh	20	11	1	6	1	6
Kera	28	6	3	1	2	6
Khuzestani Arabic	28	10	1	1	2	6
Kumiai	12	10	1	6	1	6
Kumzari	19	8	1	1	1	4
Kunama	15	10	3	6	2	4
Lizu	29	8	2	1	1	3
Lusoga	16	10	2	2	5	5
Madurese	35	8	1	1	1	3
Mah Meri	35	19	2	3	1	1
Malagasy	32	4	1	1	1	4
Malayalam	13	11	1	8	1	3
Nen	30	8	1	3	1	10
Paiute	30	11	1	5	1	4
Pitjantjatjara	22	6	1	10	1	4
Polish	39	6	1	6	3	4
Punjabi	27	17	3	2	2	3
Qanjobal	70	5	1	1	1	4
Sama	25	6	1	1	1	4
Scottish Gaelic	30	18	1	2	2	2
Seenku	29	12	4	1	1	2
Setswana	33	7	2	1	5	4
Shanghainese	18	10	5	1	1	1
Shawi	17	4	1	6	1	6
Shilluk	27	10	7	1	1	6
Shipibo	17	8	1	6	1	6
Sumi	31	6	3	6	1	4
Swedish	25	17	1	2	3	2
Telugu	17	12	1	8	3	2
Tera	31	11	3	1	1	2
Ukrainian	20	6	1	7	3	4
Urarina	28	13	2	1	1	8
Xiang	28	9	7	1	1	1
Zwara Berber	12	4	3	2	2	6
Average	**24,72**	**9,64**	**1,92**	**3,32**	**1,60**	**4,36**