The Existence of Negative Correlation Between Linguistic Measures Across Languages

Germán Coloma
(CEMA University, Buenos Aires)
(Visiting Scholar, UCSB)
Abstract
This paper proposes a procedure to evaluate the possible existence of negative correlation between three linguistic ratios (phonemes per syllable, syllables per word, and words per clause). It is based on the use of partial correlation coefficients, instrumental variables and simultaneous-equation regressions, and it is performed on data obtained from the fable “The North Wind and the Sun”, translated into 50 languages. After controlling for phenomena such as geographic factors and statistical endogeneity, we end up with the conclusion that the three ratios are negatively correlated between themselves, and this can be seen as a signal of the possible existence of complexity trade-offs.
Results from the quantitative literature

• Literature that uses “theoretical” measures: Shosted (2006) does not find any significant correlation between cross-linguistic measures of phonological and morphological complexity; Maddieson (2007) does not find anything important when comparing different phonological measures (e.g., inventory size vs. number of syllable types).

• Literature that uses “empirical” measures: Fenk-Ozclon & Fenk (2008) do find a significant negative correlation between phonemes per syllable, syllables per word and words per clause; Nettle (1995) also finds a significant negative correlation between inventory size and average number of phonemes per word.

• Nobody uses any other method but standard (Pearson) correlation coefficients.
Characteristics of this study

- It uses “empirical” measures (linguistic ratios): Phonemes per syllable, syllables per word and words per clause, for the same text (“The North Wind and the Sun”) in 50 different languages.
- Not only does it calculate standard correlation coefficients, but also partial correlation coefficients.
- It introduces different statistical sophistications (geographic factors, instrumental variables) making use of the relationship between correlation coefficients and regression coefficients.
Division by geographic areas

**Europe**: Irish, English, German, Polish, Greek, French, Hungarian, Basque, Spanish, Portuguese.

**Africa**: Tashlhiyt, Nara, Amharic, Dinka, Hausa, Temne, Kabiye, Igbo, Mono, Sandawe.

**West Asia**: Georgian, Turkish, Hebrew, Arabic, Persian, Tajik, Nepali, Hindi, Bengali, Tamil.

**East Asia**: Japanese, Korean, Mandarin, Cantonese, Thai, Burmese, Vietnamese, Malay, Tausug, Arrernte.

**America**: Sahaptin, Apache, Chickasaw, Seri, Zapotec, Trique, Quichua, Shiwilu, Yine, Mapudungun.
Illustrations of the IPA

British English: Received Pronunciation

Peter A. Ross
School of Linguistics and Applied Language Studies
University of Reading
p.ross@reading.ac.uk

The accent described here is the present-day version of the accent that has been used as the standard in phoneticians' description of the pronunciation of British English for centuries. The definition of this accent is a matter of heated debate and frequent controversy: the arguments will not be rehearsed here, but the interested reader is recommended to look at Jones (1917 and subsequent) and Wells (2000). The most important aspects of this accent should, however, be made clear:

a. The number of native speakers of this accent who originate in Ireland, Scotland, and Wales is very small and probably diminishing, and it is therefore a minority to call it an accent of British English. It is an accent spoken by some English people.

b. The great majority of native speakers of this accent are of middle-class or upper-class origin, educated at private schools and (of appropriate age) university. This does not mean that the accent cannot be acquired by others: the present author (who attended a state school in the Midlands) originally spoke with an accent with noticeable regional features, but has over many years of teaching the phonetics of English acquired an accent not far from the standard one described here.

c. The majority of speakers of this accent live in, or originate from, the south-east of England.

d. The accent is most familiar as that used by most 'official' BBC speakers of English origin (newscasters and announcers on Radio 4 and Radio 3, and most television channels). It is also frequently heard on the BBC World Service, though that service appears to have adopted the policy of sometimes using newscasters and announcers with noticeable foreign accents. It is clear that this accent will eventually lose its pre-eminent status in broadcasting as a result of the wish to broaden the social base of broadcast speech, but it will take a long time for this to happen.

The accent has been known for nearly a century as RECEIVED PRONUNCIATION, or by its abbreviation, RP. Early in the 20th century, David Jones, the great exponent of the description of English pronunciation, called it PUBLIC SCHOOL PRONUNCIATION (Jones 1917), but later changed the name to Received Pronunciation. Other names have been proposed, such as GENERAL BRITISH (GB) and EDUCATED SOUTHERN BRITISH ENGLISH. The present author's own preference is for the name BBC PRONUNCIATION or BBC ACCENT (Jones, ed. by Roach et al, 2003), but given the continuing popularity of the name Received Pronunciation, this has been used for the description which follows.

The choice of symbols for the representation of RP is one which has provoked much discussion, but since the 1980s there has existed, largely as a result of pressure from the major ELT publishers, de facto standard sets of symbolisation conventions which has remained almost unchanged to the present day. It is widely accepted that some modification of these conventions is needed to take account of observable changes in the pronunciation of English, but it is felt to be important that such modifications should not be introduced without general agreement among practitioners of English phonetics so that the benefits of a common system of transcription enjoyed over the last twenty or thirty years should not be lost. An alternative set of transcriptions has, however, been used in the Oxford Dictionary of Pronunciation (Upton et al. 2001).

Consounds

<table>
<thead>
<tr>
<th>Bilabial</th>
<th>Labiodental</th>
<th>Dorsal</th>
<th>Alveolar</th>
<th>Post-alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>b</td>
<td>t</td>
<td>d</td>
<td>k</td>
<td>g</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Nasal

<table>
<thead>
<tr>
<th>m</th>
<th>n</th>
</tr>
</thead>
</table>

Fricative

<table>
<thead>
<tr>
<th>f</th>
<th>v</th>
</tr>
</thead>
</table>

Approximant

<table>
<thead>
<tr>
<th>w</th>
</tr>
</thead>
</table>

Lateral approximant

<table>
<thead>
<tr>
<th>l</th>
</tr>
</thead>
</table>

Vowels

<table>
<thead>
<tr>
<th>o</th>
<th>a</th>
</tr>
</thead>
</table>

Plosives, with the exception of /h/ and /w/ and articulates show a distinction between VOICED and VOICELESS, or, alternatively, between LENSES and FORTIS, e.g., /b/, /v/, /d/, /z/, /g/, /j/. Voicing of so-called voiced consonants is often very weak or even undetectable, a fact which has led to the idea that force of articulation is what distinguishes pairs of obstruent consonants, /p/ for example, being classed as fortis (strongly articulated) and /b/ as lenis (weakly articulated). Vowels are markedly shortened before fortis consonants in the same syllable. The plovalve /p/, /t/, /k/ are aspirated before vowels (i.e. part of the following vowel is devoted), and in an equivalent process /b/, /d/, /w/ and /g/ are partially devoiced following these plosives. This is much less noticeable if the following vowel is unstressed. The devoicing is usually blocked if /e/ precedes the plosive within a syllable. Syllable-final /p/, /t/, /k/ are frequently preceded by a glottal stop [ʔ] unless followed by a vowel, though there are few examples of this in the recording on which our transcription is based. It sometimes happens that /ə/ is replaced, rather than preceded by, a glottal stop, particularly before a syllabic nasal,
school, and finally at Oxford University. She teaches in the English Department of a foreign university.

The original recording is on Microtrak, and has been made available to the IPA along with this analysis. The acoustic analysis of the recording was made using the NH Speech Analyzer program; the file was divided into nine chunks and phonetic labelling of acoustic segments was carried out. The phonemic transcription is idealized, and various voiced speech phenomena that are recorded in the allophonic transcription are ignored. The allophonic transcription is different from a purely phonetic transcription as described by Pike (1945) in that it adds phonetic detail to the basic phonological structure given in the phonemic transcription. It should be noted that many of the phonologically voiced consonants marked as devoiced are at times within the consonant less than fully devoiced, but transcribing both voiced and devoiced sections of them would have made the transcription too unwieldy. Pronunciation is included because information is not transcribed.

Phonemes

The North Wind said, "You will go off with a warm cloak, even if you do not want it!" The first section of medium cold made him freeze in his cloak of wood. His hands were blue, his face was red, and after a while he died. He then said, "I am going to stop."

Allophones

The North Wind was warm, and the Sun was hot. The first section of medium cold made him freeze in his cloak of wood. He then said, "I am going to stop."

Orthographic

The North Wind and the Sun were disputing which was the stronger. When a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

References

Orthographic version
The North Wind and the Sun were disputing which was the stronger, when a traveller came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveller take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveller fold his cloak around him, and at last the North Wind gave up the attempt. Then the Sun shone out warmly, and immediately the traveller took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

Phonemic version

Clauses = 9 ; Words = 113 ; Syllables = 143 ; Phonemes = 383 .
Phonemes/Syllables = 2.6783 ; Syllables/Words = 1.2655 ; Words/Clause = 12.56
<table>
<thead>
<tr>
<th>Concept</th>
<th>Minimum</th>
<th>Average</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phonemes/Syllables</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>2.0430</td>
<td>2.3128</td>
<td>2.7089</td>
</tr>
<tr>
<td>Africa</td>
<td>1.7115</td>
<td>2.1400</td>
<td>2.5521</td>
</tr>
<tr>
<td>West Asia</td>
<td>2.1468</td>
<td>2.3079</td>
<td>2.5288</td>
</tr>
<tr>
<td>East Asia</td>
<td>1.9559</td>
<td>2.4525</td>
<td>2.8547</td>
</tr>
<tr>
<td>America</td>
<td>1.9617</td>
<td>2.3028</td>
<td>2.5761</td>
</tr>
<tr>
<td><strong>Syllables/Words</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>1.2655</td>
<td>1.7569</td>
<td>2.2530</td>
</tr>
<tr>
<td>Africa</td>
<td>1.5739</td>
<td>1.9910</td>
<td>2.7553</td>
</tr>
<tr>
<td>West Asia</td>
<td>1.6640</td>
<td>2.3937</td>
<td>3.1899</td>
</tr>
<tr>
<td>East Asia</td>
<td>1.0000</td>
<td>2.1201</td>
<td>3.1190</td>
</tr>
<tr>
<td>America</td>
<td>1.5414</td>
<td>2.5112</td>
<td>3.7460</td>
</tr>
<tr>
<td><strong>Words/Clauses</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>9.89</td>
<td>12.12</td>
<td>18.43</td>
</tr>
<tr>
<td>Africa</td>
<td>8.44</td>
<td>11.27</td>
<td>13.83</td>
</tr>
<tr>
<td>West Asia</td>
<td>7.33</td>
<td>10.10</td>
<td>15.63</td>
</tr>
<tr>
<td>East Asia</td>
<td>6.00</td>
<td>9.73</td>
<td>16.71</td>
</tr>
<tr>
<td>America</td>
<td>5.70</td>
<td>8.62</td>
<td>14.27</td>
</tr>
</tbody>
</table>
SylWord = -0.4327 * PhonSyl + 3.1513
R² = 0.0311

- Yine
- Tamil
- Chickasaw
- Japanese
- Amharic
- Igbo
- Spanish
- Mandarin
- Mono
- Irish
- English
- Vietnamese
WordClause = -0.9142 * PhonSyl + 12.474

$R^2 = 0.0059$
Standard correlation coefficient

- Formula: \[ r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \]

- Example for the correlation between syllables per word and words per clause:

\[ r = \frac{\sum (x_i - 2.1546)(y_i - 10.37)}{\sqrt{\sum (x_i - 2.1546)^2 \cdot \sum (y_i - 10.37)^2}} = -0.6806 \]
Correlation coefficients in the analyzed database

<table>
<thead>
<tr>
<th>Variable</th>
<th>Phoneme/Syllable</th>
<th>Syllable/Word</th>
<th>Word/Clause</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemes per syllable</td>
<td>1.0000</td>
<td>-0.1763</td>
<td>-0.0831</td>
</tr>
<tr>
<td>Syllables per word</td>
<td></td>
<td>1.0000</td>
<td>-0.6806</td>
</tr>
<tr>
<td>Words per clause</td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>

$t$-statistic:
$$t = \frac{r \cdot \sqrt{d}}{\sqrt{1 - r^2}} = \frac{r \cdot \sqrt{n - 2}}{\sqrt{1 - r^2}} = \frac{r \cdot \sqrt{48}}{\sqrt{1 - r^2}}$$

$p$-value:
$$p = 1 - F(t)$$
Correlation calculated using results from a statistical regression

- Example of regression between syllables per word and words per clause:
  \[ WC = 16.941 - 3.0507 \cdot SW \quad R^2 = 0.4632 \]
  \[ SW = 3.7288 - 0.1518 \cdot WC \quad R^2 = 0.4632 \]

- **Method 1:**
  \[ r = -\sqrt{R^2} = -\sqrt{0.4632} = -0.6806 \]
Correlation calculated using results from a statistical regression (cont.)

• **Method 2:**

\[
r = -\sqrt{\beta_x \cdot \beta_y} = -\sqrt{(-3.0507) \cdot (-0.1518)} = -0.6806
\]

(This method is also good for multiple regression equations, in which we obtain **partial correlation coefficients**.)
Results from the regression of a system of equations (OLS)

\[ PS = 3.0408 - 0.1768 \cdot SW - 0.0344 \cdot WC \]
\[ SW = 5.0995 - 0.5755 \cdot PS - 0.1562 \cdot WC \]
\[ WC = 22.610 - 2.3061 \cdot PS - 3.2164 \cdot SW \]

- Partial correlation coefficients:

\[ r(PS, SW) = -\sqrt{(-0.1768) \cdot (-0.5755)} = -0.3190 \]
\[ r(SW, WC) = -\sqrt{(-0.1562) \cdot (-3.2164)} = -0.7088 \]
\[ r(PS, WC) = -\sqrt{(-0.0344) \cdot (-2.3061)} = -0.2817 \]
Inclusion of other (geographic) variables in the system

\[ PS = c(1) \times \text{Europe} + c(2) \times \text{Africa} + c(3) \times \text{Westasia} + c(4) \times \text{Eastasia} + c(5) \times \text{America} + c(6) \times \text{SW} + c(7) \times \text{WC} \]

\[ \text{SW} = c(11) \times \text{Europe} + c(12) \times \text{Africa} + c(13) \times \text{Westasia} + c(14) \times \text{Eastasia} + c(15) \times \text{America} + c(16) \times \text{PS} + c(17) \times \text{WC} \]

\[ \text{WC} = c(21) \times \text{Europe} + c(22) \times \text{Africa} + c(23) \times \text{Westasia} + c(24) \times \text{Eastasia} + c(25) \times \text{America} + c(26) \times \text{PS} + c(27) \times \text{SW} \]
Seemingly unrelated regressions (SUR)

• Statistical procedure in which several equations are run at the same time, making use of the information gathered when estimating each equation to improve the estimation of the other equations.

• It consists of two consecutive stages. On the first, one runs the equations using OLS. Then, one calculates the correlations between the equations’ residuals.

• Using those correlations, one re-runs the equations and finds improved regression coefficients.
Partial correlation coefficients for the system with geographic variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Phon/Syl</th>
<th>Syl/Word</th>
<th>Word/Claus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phonemes per syllable</td>
<td>1.0000</td>
<td>-0.6003</td>
<td>-0.4896</td>
</tr>
<tr>
<td>Syllables per word</td>
<td></td>
<td>1.0000</td>
<td>-0.9120</td>
</tr>
<tr>
<td>Words per clause</td>
<td></td>
<td></td>
<td>1.0000</td>
</tr>
</tbody>
</table>
Endogeneity problem

- It is a problem that arises when one uses explanatory variables that are at the same time explained in the system of equations.
- None of them is truly determined by the other ones, but all of them are simultaneously defined through a process governed by a pre-established environment.
- To deal with this, one uses instrumental variables (i.e., variables related with the endogenous variables under analysis, but exogenously determined).
Use of instrumental variables

Phonological: Number of consonants, number of vowels, number of distinctive tones, stress distinctiveness, syllable complexity (Source: IPA).

Morphological: Number of noun genders, number of noun cases, number of verbal inflections, type of inflection (Source: WALS).

Syntactic: Order of object and verb, order of adjective and noun, morphosyntactic alignment (Source: WALS).

Other: Genetic affiliation (Indo-European or not).
Welcome to WALS Online

The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors (http://wals.info/author).

The first version of WALS was published as a book with CD-ROM in 2005 by Oxford University Press (http://ilclub.linguateca.mapress.com/product/9781845205817.do). The first online version was published in April 2008. The second online version was published in April 2011.

The 2013 edition of WALS corrects a number of coding errors especially in Chapters 1 and 3. A full list of changes is available here (http://wals.info/changes).

Starting with this edition of WALS, there will not be specific editions every two or three years, but we will be updating it whenever corrections or additions are made. Changes in value assignment will be made transparent by showing a history on the respective pages.

WALS Online is a publication of the Max Planck Institute for Evolutionary Anthropology (http://www.eva.mpg.de). It is a separate publication, edited by Dryer, Matthew S., & Haspelmath, Martin. (Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013). The main programmer is Robert Forkel.

How to use WALS Online

Using WALS Online requires a browser with Javascript enabled.

WALS News

WALS 2013

Thu, 14 Mar 2013 05:20:35 -0600

What’s new? Updated data, better integration of sources, the web application is now based on the old framework ...

Lateste Comments

Comment on Datapoint Kukur / Order of Subject and Verb by Thomas Hart Chappell

http://blog.wals.info/datapoint/82a-wals_code_kk/

Thu, 25 Mar 2014 11:33:11 -0700

I guess I posted the same question twice. I apologize.

Interactive Reference Tool (WALS program)


Terms of use

The content of this website is published under a Creative Commons Licence. We invite the community of users to think about further applications for the
Three-stage least squares (3SLS)

- Statistical procedure in which the endogenous variables are replaced by (exogenous) instrumental variables.
- It consists of three consecutive stages. On the first, one runs OLS equations for the endogenous variables against all the instrumental variables, and creates new variables using the “fitted values” from the regressions.
- Using those new variables, one runs the desired regression equations and then calculates the correlations between the statistical errors of the equations (second stage).
- Using those correlations, one then re-runs the equations and finds improved regression coefficients (third stage).
<table>
<thead>
<tr>
<th>Concept</th>
<th>Phon/Syl vs Syl/Word</th>
<th>Syl/Word vs Word/Clause</th>
<th>Phon/Syl vs Word/Clause</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard correlation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient</td>
<td>-0.1763</td>
<td>-0.6806</td>
<td>-0.0831</td>
</tr>
<tr>
<td>Probability</td>
<td>(0.2206)</td>
<td>(0.0000)</td>
<td>(0.5661)</td>
</tr>
<tr>
<td><strong>Partial correlation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient</td>
<td>-0.3190</td>
<td>-0.7088</td>
<td>-0.2817</td>
</tr>
<tr>
<td>Probability</td>
<td>(0.0255)</td>
<td>(0.0000)</td>
<td>(0.0499)</td>
</tr>
<tr>
<td><strong>Geogr. variables (SUR)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient</td>
<td>-0.6003</td>
<td>-0.9120</td>
<td>-0.4896</td>
</tr>
<tr>
<td>Probability</td>
<td>(0.0000)</td>
<td>(0.0000)</td>
<td>(0.0006)</td>
</tr>
<tr>
<td><strong>Geogr. variables (3 SLS)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coefficient</td>
<td>-0.4933</td>
<td>-0.9745</td>
<td>-0.3043</td>
</tr>
<tr>
<td>Probability</td>
<td>(0.0006)</td>
<td>(0.0000)</td>
<td>(0.0422)</td>
</tr>
</tbody>
</table>

Not Significant at 10%

Significant at 1%

Significant at 5%
Conclusions

• Syllables per word are negatively correlated with words per clause, whichever method one uses.
• Using partial correlations (and even more using SUR and 3SLS) phonemes per syllable are negatively correlated with syllables per word.
• Negative correlation between phonemes per syllable and words per clause is also highly significant if we use SUR, (and is significant as well using 3SLS).
• All this can be interpreted as a signal for the possible existence of “complexity trade-offs”.