Japan Bilingual Publishing Co.

Linguistic Exploration

https://ojs.bilpub.com/index.php/le

ARTICLE

An Extended Synergetic Model of Language Phonology

Germán Coloma

Department of Economics, CEMA University, Buenos Aires C1054AAP, Argentina

ABSTRACT

This paper extends the analysis originally performed by the author in 2014, by developing a model based on the principles of the so-called "synergetic linguistics approach". This model tries to explain the occurrence of several phonological characteristics of languages as a process of maximization of a welfare function, which considers both the ease of decoding language expressions and the effort to produce those expressions. The main changes in this paper are the use of a larger and more balanced sample of 150 languages, the inclusion of new phonological variables, and the explicit consideration of phylogenetic, geographic and demographic factors. The analysis is carried out using seemingly unrelated regressions for a system of equations that relate six characteristics of languages: (1) number of consonant phonemes, (2) vowel qualities, (3) distinctive tones, (4) use of stress, (5) vowel length and (6) vowel nasalization, identifying those equations as first-order conditions in a welfare maximization problem. The main finding is that the key phonological variable seems to be the number of vowel qualities in a language, which is positively correlated with the number of consonants and the use of vowel length, and negatively correlated with vowel nasalization. Other important determinants seem to be the use of contrasting vowel length, and the existence of stress distinctions.

Keywords: Synergetic Linguistics; Welfare Maximization; Phonological Variables; Seemingly Unrelated Regressions; Correlation

*CORRESPONDING AUTHOR:
Germán Coloma, Department of Economics, CEMA University, Buenos Aires C1054AAP, Argentina; Email: gcoloma@ucema.edu.ar

ARTICLE INFO
Received: 16 November 2024 | Revised: 30 November 2024 | Accepted: 5 December 2024 | Published Online: 10 December 2024
DOI: https://doi.org/10.55121/le.v1i1.192

CITATION
Coloma, G., 2024. An Extended Synergetic Model of Language Phonology. Linguistic Exploration. 1(1): 12–35.
DOI: https://doi.org/10.55121/le.v1i1.192

1. Introduction

In 2014, the author presented a statistical model that related the number of consonant and vowel phonemes that languages have, together with other variables that measured the use of tone and stress as means to distinguish between different meanings of otherwise identical expressions[1]. This model was based on the principles of the so-called "synergetic linguistics approach", which implies analyzing the functioning of language using elements taken from the general theory of systems.

The approach taken proved to be fruitful, in the sense that the author obtained a series of conclusions about the relationships between several phonological characteristics of languages, assuming that those characteristics were chosen in order to maximize a certain "welfare function". The main conclusion pointed out towards the central role of the stress variable, in the sense that languages in which stress is distinctive or non-predictable tend to have fewer consonants and vowels (and are also less likely to use tone as an additional phonological device).

That conclusion, however, is an empirical one, and depends heavily on the data used in the abovementioned paper[1]. That data consisted of a series of observations on different languages, chosen from various families and geographic locations. The sample as a whole, however, was strongly biased towards "major languages" (i.e., languages with several million native speakers) and towards languages from Europe and Asia. One of the main ways to extend the proposed model is therefore to increase the number of languages in our sample, and to make that sample more diverse in both geographic and phylogenetic aspects.

That is what is conducted in this paper, in which an extended synergetic model of language phonology is developed which has more phonological variables, related to the number of distinctive tones that languages have, and to the possible use of vowel nasalization and vowel length as phonemic devices. Additionally, the model incorporates some phylogenetic, geographic and demographic dimensions, not only through the selection of languages, but also as exogenous variables that may influence the choice of the phonological characteristics of those languages. All this is done using the same approach applied in our previous work, which implies the use of a specific statistical method, designed for the simultaneous regression of equation systems[1].

The paper is organized as follows. Section 2 will present the theoretical model, while Section 3 will describe the data that we use. Section 4, in turn, will be devoted to the main empirical results obtained, using both the original model and the extended version of the model. Finally, Section 5 will make a few concluding remarks about the whole paper.

2. The Model

The synergetic model that the author used in 2014 to interpret the relationships between different phonological variables implied the maximization of a certain welfare function for languages, based on two main considerations: the decoding ease of those languages, and their corresponding production effort[1]. The idea behind this is that language (and the sub-systems that are part of it) can be seen as a self-organizing and self-regulating system whose properties come from the interaction of several constitutive, forming and control requirements. See, for example, Kohler[2] or Klymenko and Yenikeyeva[3].

In the case of the model that we apply, we interpret that the most relevant requirements for language use are the idea that it must be good to communicate concepts through expressions that have meanings, and that it must possess as few elements as possible to produce the desired expressions. The first of those properties is therefore related to the process of coding and decoding expressions (for example, through words that have different meanings), while the second one has to do with the production of the corresponding expressions (for example, through sounds that combine according to certain rules).

The different languages that are spoken in the world cope with those requirements in very different ways, but they all share the common characteristic of using "combinatorial phonology". This implies choosing a (relatively small) set of sounds, defined in a certain way, and combining the elements of that set in order to produce words with different meanings. For example, in English we have the word "tap", which is made up of three sounds, that can be written phonetically as /t/, /æ/ and /p/. However, the same three sounds, in different order, can be combined to produce other words such as "pat" and "apt". This means that /t/, /æ/ and /p/ are different "phonemes", whose combination can generate words with different meanings.

All spoken languages follow a similar strategy, in the sense that they define a number of specific phonemes that they combine to produce words. Something similar occurs with sign languages, which usually employ combinations of signs rather than isolated self-meaning units (see, for example, Brentari[4]).

The sounds that spoken languages use, however, are not always the same, and it is very easy to find some phonemes that exist in English but not in other languages (for example, /θ/, which is the first sound of the word "thing", and does not exist in the majority of languages) and some phonemes that are used in other languages but not in English (for example, /ɲ/, which is the second sound of the Spanish word "*año*", that means "year"). All spoken languages, however, seem to use at least some "vowel phonemes" and some "consonant phonemes", since there is no known language that uses only vowels or only consonants. All these similarities, and others concerning the syntactic structure of languages, point out in the direction of the idea that spoken language was "invented only once", around 70,000 to 80,000 years ago, and then it started to diverge as human beings began to spread throughout the world. For more information about this, see the work by Hurford[5].

The size of the phoneme inventories of the different languages is very variable, as is the division of those inventories between vowels and consonants. The Quechua language, for example, has only three vowels (/i/, /a/ and /u/), but English has eleven (/i/, /ɪ/, /e/, /æ/, /ə/, /ʌ/, /ɑ/, /ɒ/, /ɔ/, /ʊ/ and /u/). Additionally, while the number of consonant phonemes in Finnish is only 13, the Lithuanian language has 45 consonant phonemes.

Languages also differ in the use of some characteristics that may change the pronunciation of the different phonemes, especially the vowels. Two of the most widely used are stress and tone. In English, for example, stress is variable, since there are words whose stressed syllable is the first one (e.g., "object") while other words carry their stress in their second syllable (e.g., "objective") or third syllable (e.g., "objectivity"). However, English words do not change their meaning due to differences in tone, while many other languages (e.g., Mandarin Chinese) have several distinctive tones (e.g., high, low, rising, falling). In Mandarin, for example, the word "*ba*" means "eight" (high tone), "to hold" (low tone), "to pull out" (rising tone) or "father" (falling tone).

Our synergetic model of language phonology tries to interpret the phonological characteristics of the different languages, assuming that each of them chooses a combination that is optimal to maximize the difference between a "decoding ease function" (D) and a "production effort function" (P). Decoding ease can in general be seen as positively related to the existence of many distinctions in the set of phonological variables. If, for example, a language had only one consonant and one vowel (e.g., /p/ and /a/), then its only possible words would be utterances such as "papa", "apap", "pappap", etc. With that limited set of options, it would be very difficult to develop a useful vocabulary.

However, having too many distinctions (e.g., more than 100 consonants, 30 vowels, or 10 distinctive tones) would probably imply a very high burden in terms of producing the different utterances needed to speak a language, and in terms of learning the language itself. So the balance between decoding ease and production effort, expressed through a "welfare function" (W), would probably imply choosing more moderate values for the different phonological variables (Consonants, Vowels, Tones, Stress, etc.), and combining them in a particular way.

In 2014, the author proposed a model in which languages are supposed to maximize a linear-quadratic welfare function similar to this:

$$\begin{gathered} W = D(X_1, X_2, \ldots, X_n) - P(X_1, X_2, \ldots, X_n) = \\ \sum_i a_i \cdot X_i + \tfrac{1}{2} \sum_i \sum_{j \neq i} a_{ij} \cdot X_i \cdot X_j - \tfrac{1}{2} \sum_i b_i \cdot X_i^2 \end{gathered} \tag{1}$$

where X_1, X_2, ..., X_n are different phonological variables (e.g., number of consonants, number of vowels, number of tones, etc.), and a_i, a_{ij} and b_i are parameters[1]. It is also assumed that, for any pair of languages "i" and "j", it holds that "$a_{ij} = a_{ji}$", so the number of actual parameters is considerably reduced. Maximizing this function implies fulfilling "n" different first-order conditions of the following form:

$$\begin{gathered} \frac{\partial W}{\partial X_i} = a_i + \sum_{j \neq i} a_{ij} \cdot X_j - b_i \cdot X_i = 0 \\ \rightarrow X_i = \frac{a_i}{b_i} - \sum_{j \neq i} \frac{a_{ij}}{b_i} \cdot X_j \end{gathered} \tag{2}$$

and that implies choosing each individual variable (X_i) as a function of the other variables (X_j). For a complete explanation of the logic behind this type of maximization, see the work by Sundaran[6], chapter 2.

The empirical strategy to find the implicit parameters of the welfare function consists of running a series of regressions between the observed values of the phonological variables for different languages, one for each of the first-order conditions implied by Equation (2). For example, the author built a database of 100 observations (each of them corresponding to a different language) with information about number of consonant phonemes (C), number of vowel phonemes (V), use of tone as a distinctive feature (T), and use of distinctive or non-predictable stress (S)[1]. These last two variables are "categorical" or "binary", in the sense that they can only have two values: 1 (if tone or stress is distinctive) and 0 (if it is not).

With that information, a system of four regression equations was run that allowed us to estimate values for the parameters of the welfare function, and also gave us a hint about the interaction of the phonological variables inside the system. The result that we got was that C, V and T were not related among themselves, but that they were all (negatively) correlated to S. This result is represented in **Figure 1**.

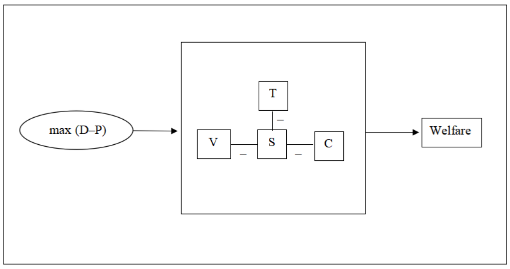

Figure 1. Diagram of the model developed by Coloma[1].

The logic behind this model is that the phonological variables that interact in the system produce certain results that are aimed at maximizing the difference between decoding ease and production effort. This generates a certain level of welfare due to the existence of some specific welfare function parameters, which are discovered through the regression analysis. This model, however, could be more complete if we allowed for the interplay between the phonological variables (chosen by the language system) and other non-linguistic variables that may also influence the choice of the different language characteristics.

The three types of non-linguistic variables that are easier to introduce in this context have to do with phylogenetic factors, geographic factors and demographic factors. Phylogenetic factors imply the existence of relationships between languages through "common ancestors". For example, Spanish and Italian are relatively similar because they both share an immediate ancestor (i.e., the Latin language), and that is why they are both said to belong to the same "language *genus*" (i.e., the group of Romance languages). However, Latin also has a common ancestor with the ancient Germanic language from which modern English is descended, so English, Spanish and Italian all belong to the same "language family" (i.e., the Indo-European family), together with many other languages such as French, German, Greek, Irish, Russian, Persian, Hindi, etc. This categorization of languages into families and *genera* is basically the one used in the work of Dryer and Haspelmath[7].

Conversely, languages such as Mandarin, Arabic, Japanese, Filipino, Yoruba, Hungarian, Cherokee and many others belong to other families, different from the Indo-European one (and different from each other's families, too). Therefore, belonging to a certain language family could be a factor that may influence the choice of phonemes and other phonological variables, which is not actually determined by the language system itself but is exogenous to that system.

A similar situation occurs with the emergence of languages in certain continents or regions (geographic factors). Languages could then be classified according to the areas where they originated, and also according to the expansion that they achieved. This last phenomenon can be approximated by the number of native speakers for each language, which is the main demographic factor related to it. This can be used to create categories such as "major language", "medium-sized language", "minor language", etc.

With the inclusion of exogenous variables related to phylogenetic (Phylog), geographic (Geogr) and demographic factors (Demog), **Figure 1** is modified and presented as **Figure 2**. **Figure 2** also includes new phonological variables in the system and, of course, we could collect data on more languages and on languages from more additional families and/or geographic locations. That is what will be conducted in the following sections of this paper, trying to see if the conclusions about the relationships between the main variables that describe the different languages' phonological systems, that the author obtained in 2014, change substantially when we include other languages, other phonological variables, and some non-linguistic variables.

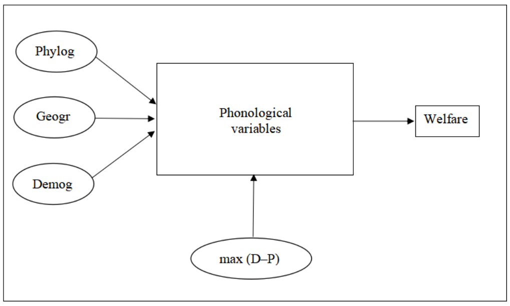

Figure 2. Modified diagram for the extended synergetic model.

3. Description of the Data

As already mentioned, the author tested the model presented in 2014 using a database of 100 languages that was specifically built for the occasion[1]. That database had a number of problems concerning its representativeness, since its main concern was to include all the languages that had more than 40 million native speakers. Because of that, it was heavily biased towards a few language families (basically, towards the Indo-European family), and it included many languages that were very similar from a linguistic point of view (e.g., Portuguese, Spanish, Italian and French; or Bengali, Hindi, Marathi and Punjabi).

One of the main changes introduced in this paper is that the database has been completely rebuilt, which now has 150 languages from a much larger set of families. Despite still having several Indo-European languages, each of them now belongs to a different *genus*, and the same occurs with all the other language families that have more than one language in the sample. The represented *genera* for the Indo-European language family are the following: Albanic (Albanian), Armenic (Armenian), Baltic (Lithuanian), Celtic (Irish), Germanic (English), Hellenic (Greek), Indic (Hindi), Iranian (Persian), Romance (Spanish) and Slavic (Russian).

Consequently, this paper ends up with a database representing 55 different language families and 148 different *genera*. The language family that has more observations is now the Niger-Congo family (with 11 languages), and there are also two creole languages in the database (Haitian and Tok-Pisin).

This new sample is also much more diverse in terms of geographic representativeness. It includes 23 North American languages, 17 South American languages, 11 European languages, 31 African languages, 44 Asian languages and 24 languages from Australasia. The author still tried to use languages that were relatively important in order to represent each *genus* included in the database, and that is why we avoided languages with fewer than one thousand native speakers. This limitation was not a problem to select languages from the main families of Europe, Africa and Asia, but it was a considerably challenging task for the languages of Australasia and the Americas.

Another requisite used to build the sample was the availability of relatively modern phonological descriptions of the included languages. Whenever possible, the author referred to the illustrations from the work by IPA[8] or articles published since 1999 in the *Journal of the International Phonetic Association*. When such sources were unavailable, the author relied on published grammar, and, in many cases, Ph.D. dissertations devoted to specific language descriptions.

As a consequence of all this, this paper ended up with the languages whose list appears in **Appendix A**, which are also shown on the map of **Figure 3**. Notice that some regions are very rich in terms of languages from different families in a relatively small area. This is, for example, the case of Mesoamerica, the Caucasus, the Gulf of Guinea, the island of New Guinea, and the Southeastern part of Asia. Conversely, some other regions such as Siberia, the Sahara and the Southern part of South America are large areas with very few autochthonous languages.

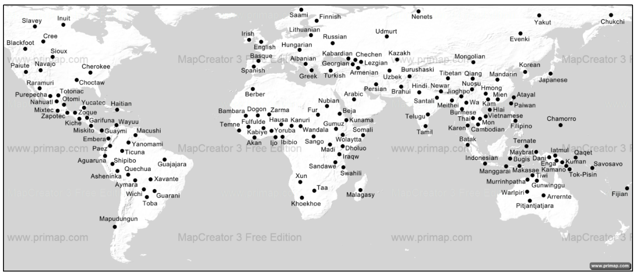

Figure 3. Languages included in the database.

Our new database is also larger (than the one developed in 2014) in terms of the included linguistic variables[1]. In addition to the data about the number of consonant and vowel phonemes, and the categorical variable concerning stress distinctiveness, for the present article we have redefined the tone variable, using the number of distinctive tones that each language has (instead of a binary variable for tonal vs. non-tonal languages). This implies that non-tonal languages (e.g., English, Spanish, Swahili) all have 1 tone each, but tonal languages may have any number of tones, from a minimum of 2 (e.g., Japanese, Navajo, Hausa) to a maximum that, in our sample, is equal to 10 (and corresponds to the Brazilian language Ticuna, and to the Chinese language Kam).

Besides, variables related to the number of vowel qualities, the use of vowel length, and the use of vowel nasalization have also been included. All of them have certain relationships with the number of vowel phonemes in the different languages, which can be illustrated with a few examples. For instance, Arabic and Indonesian are languages that both have six vowel phonemes, while in Indonesian each phoneme represents a different "sound quality" (/i/, /e/, /ə/, /a/, /o/ and /u/), in Standard Arabic there are only three qualities (/i/, /a/ and /u/) and the other three vowel phonemes are the long counterparts of those basic sounds (i.e., /i:/, /a:/ and /u:/). The Guarani language, in turn, also has six vowels qualities, just like Indonesian (in this case, they are /i/, /ɨ/, /e/, /a/, /o/ and /u/), but it also has an additional set of nasalized vowels (/ĩ/, /ɨ/, /ẽ/, /ã/, /õ/ and /ũ/) that makes their total number of vowel phonemes equal to 12. We can also mention the case of the Navajo language, which has only four vowel qualities (/i/, /e/, /a/ and /o/) but 16 vowel phonemes. This is because it has both oral and nasal vowels (/ĩ/, /ẽ/, /ã/ and /õ/), and also long vowels (/i:/, /ĩ:/, /e:/, /ẽ:/, /a:/, /ã:/, /o:/ and /õ:/).

The main statistics concerning our 150-language database are summarized in **Table 1**, and the complete set of values corresponding to the 150 languages are in **Appendix B**. **Table 1** has classified languages according to the region in which they are spoken, defining eight large macro-areas: North America, South America, East Africa, West Africa, Europe, West Asia, East Asia and Australasia. Separate information for the languages that belong to the seven largest language families (Niger-Congo, Indo-European, Austronesian, Sino-Tibetan, Afro-Asiatic, Altaic and Nilo-Saharan) is also provided. Languages have also been classified in categories considering their "sizes" (i.e., their number of native speakers). Major ones are those with more than 10 million speakers, and that group is represented by 29 languages in the used sample. Conversely, languages with fewer than 100 thousand speakers are considered to be "minor", and that is the case for 43 languages in the database. The remaining 78 languages are therefore "medium-size languages" (i.e., between 0.1 and 10 million native speakers).

Table 1 shows that the average number of consonant phonemes in the whole sample is equal to 24.39, but that number varies a lot by group of languages. The languages from Australasia, for example, have an average of 17.46 consonants, while the East African languages have an average of 35.71 consonants. It can also be observed that, while the average number of vowel phonemes is 9.21, the average number of vowel qualities (VowQual) is 6.11, and that the ranking of those variables is different. For example, the area with the largest vowel phoneme average is East Africa, while the area with the largest vowel quality average is East Asia.

The number of tones that languages have is also very variable among regions. It can be seen, for example, that the European languages included in our sample have only one distinctive tone (i.e., they are non-tonal languages), but the average East Asian language has 3.21 distinctive tones (**Table 1**). On the other extreme, none of the included West African languages make distinctions based on the use of stress, while 54.5% of the European languages do so.

Finally, there are also significant differences across language families. For example, the average Altaic language has 12.38 vowel phonemes, and the average Austronesian language has only six vowel phonemes. Similarly, while 60% of the included Indo-European languages have distinctive or non-predictable stress, none of the eight included Nilo-Saharan languages possesses that characteristic.

In our 150-language database, the number of consonant phonemes in each language seems to be positively correlated with the number of vowel phonemes. Their direct coefficient of correlation is equal to 0.2896, and this can be somewhat observed in **Figure 4**, where each language is depicted as a point in the consonant vs. vowel space (and there is a positively-sloped line that represents correlation).

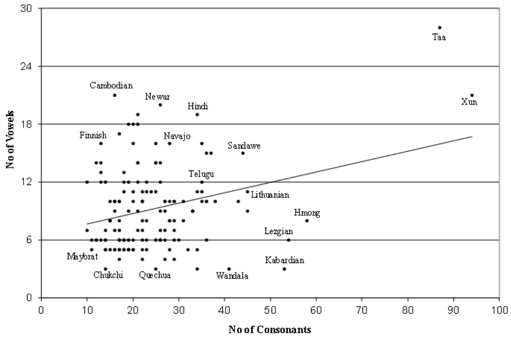

Figure 4. Number of consonant and vowel phonemes.

Figure 4 demonstrates that most languages are grouped

Table 1. Average values of the phonological variables.

Concept	Consonants	Vowels	VowQual	Tones	Stress	%
North America	19.78	8.91	5.04	1.78	39.1%	15.3%
South America	19.18	9.00	5.35	1.53	41.2%	11.3%
East Africa	35.71	10.94	6.00	2.29	5.9%	11.3%
West Africa	23.43	10.43	6.79	2.21	0.0%	9.3%
Europe	27.36	9.18	7.36	1.00	54.5%	7.3%
West Asia	29.80	9.10	6.15	1.05	30.0%	13.3%
East Asia	26.13	10.75	7.67	3.21	12.5%	16.0%
Australasia	17.46	6.25	5.17	1.21	50.0%	16.0%
Niger-Congo	21.91	11.27	7.64	2.18	0.0%	7.3%
Indo-European	29.20	8.70	7.90	1.00	60.0%	6.7%
Austronesian	20.20	6.00	5.50	1.00	50.0%	6.7%
Sino-Tibetan	29.78	11.00	7.44	2.78	0.0%	6.0%
Afro-Asiatic	29.13	7.38	4.88	1.75	12.5%	5.3%
Altaic	21.00	12.38	8.13	1.13	12.5%	5.3%
Nilo-Saharan	25.63	9.88	6.63	2.50	0.0%	5.3%
Major (> 10 million)	23.41	9.69	6.97	1.79	24.1%	19.3%
Medium-size	25.40	9.27	6.40	2.00	25.6%	52.0%
Minor (< 0.1 million)	23.23	8.77	5.00	1.56	39.5%	28.7%
Total Sample	24.39	9.21	6.11	1.83	29.3%	100.0%

together in a space that ranges between 5 and 11 vowels, and between 14 and 30 consonants. Some other languages, however, are outside that range, and the two most notable outliers are the Khoisan languages Taa (spoken in Botswana) and Xun (spoken in Angola), which have more than 80 consonant phonemes and more than 20 vowel phonemes each. This is heavily influenced by the fact that those languages possess many vowel distinctions based on length and nasalization, and they also have many complex consonants such as clicks and glottalized sounds. Other languages are noticeable because of their extreme imbalance between vowels and consonants. For example, the Kabardian language (West Caucasian, spoken in Russia) has 53 consonants and only three vowel phonemes. On the other extreme of the spectrum, the Cambodian language has more vowel phonemes than consonant phonemes (21 vs. 16).

Another possible correlation of interest is the one between consonant phonemes and vowel qualities. Their correlation coefficient is also positive (r = 0.0513), but it is much smaller than the one calculated for consonant phonemes versus vowel phonemes. That is why the correlation line depicted in **Figure 5** is flatter than the one depicted in **Figure 4**. However, **Figure 5** also shows a high concentration of languages in a relatively small number space, which in this case is the range between four and eight vowel qualities and between 14 and 30 consonant phonemes.

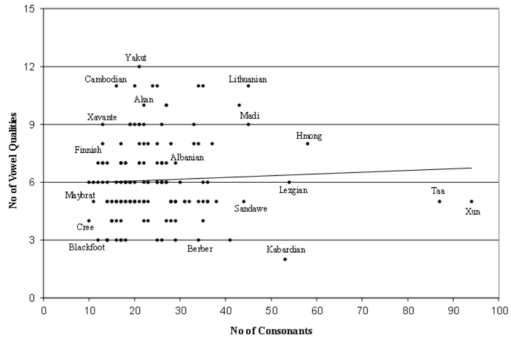

Figure 5. Consonant phonemes vs. vowel qualities.

All the other correlations between phonological variables also show relatively small coefficients. **Table 2** shows that eight of these coefficients turn out to be negative, while the remaining seven are positive. However, their maximum absolute value is below 0.31, which may imply that, as found in 2014, the true relationships between the variables under analysis are indirect[1].

The empirical procedure that we will use to unravel those indirect relationships between the phonological variables in our sample of languages will be based on the model that we described in Section 2, and it will consist of a series of statistical regressions. That is what will be performed in the following sections, in which the original model devel-

Table 2. Correlation coefficients for the phonological variables.

Concept	Consonants	VowQual	Tones	Stress	VowNasal	VowLength
Consonants	1.0000					
Vowel Qualities	0.0513	1.0000				
Tones	0.1374	0.1726	1.0000			
Stress	−0.0973	−0.1033	−0.3031	1.0000		
Vowel Nasalization	0.1528	0.0641	0.0605	−0.0188	1.0000	
Vowel Length	0.0818	−0.2445	−0.0070	−0.2055	−0.0616	1.0000

oped by the author will be tested against the newly assembled database. The work will also extend the model to include new phonological variables and some possible connections with several non-linguistic variables.

4. Empirical Results

4.1. Application to the Original Model

The first logical step in analyzing our new sample of languages is to use its data to replicate the results of the synergetic phonology model developed in 2014[1]. As already mentioned, that model consists of a system of four equations whose variables are the number of consonant phonemes (Consonants), the number of vowel phonemes (Vowels), a categorical variable that takes a value equal to one when a language has distinctive tones (Tonal), and another categorical variable that takes a value equal to one when stress is distinctive or non-predictable (Stress), as written in

$$\text{Consonants} = \text{c}(1) + \text{c}(2) * \text{Vowels} + \text{c}(3) * \text{Tonal} + \text{c}(4) * \text{Stress} \tag{3}$$

$$\text{Vowels} = \text{c}(11) + \text{c}(12) * \text{Consonants} + \text{c}(13) * \text{Tonal} + \text{c}(14) * \text{Stress} \tag{4}$$

$$\text{Tonal} = \text{c}(21) + \text{c}(22) * \text{Consonants} + \text{c}(23) * \text{Vowels} + \text{c}(24) * \text{Stress} \tag{5}$$

$$\text{Stress} = \text{c}(31) + \text{c}(32) * \text{Consonants} + \text{c}(33) * \text{Vowels} + \text{c}(34) * \text{Tonal} \tag{6}$$

Through which, a series of regressions (Regression 1) was run using the same statistical method applied in 2014[1]. This method is known as “seemingly unrelated regressions” (SUR), a variation of the more traditional “ordinary least-squares method,” specifically designed for regressing systems of equations. For an explanation about the logic behind this method, see the work by Greene[9], chapter 10.

The results obtained from this estimation (Regression 1) are presented in the first two columns of **Table 3**. As observed, several coefficients in this regression are statistically insignificant. Therefore, a new regression (Regression 2) was conducted, excluding those coefficients, until all estimated coefficients were statistically significant at the 5% probability level. These regressions, as all the others whose results are reported in this paper, were performed using the software package EViews 10.

The results of the estimations under Regression 2 imply that the analysis leads to the following system:

$$\text{Consonants} = \text{c}(1) + \text{c}(2) * \text{Vowels} + \text{c}(3) * \text{Tonal} \tag{7}$$

$$\text{Vowels} = \text{c}(11) + \text{c}(12) * \text{Consonants} + \text{c}(14) * \text{Stress} \tag{8}$$

$$\text{Tonal} = \text{c}(21) + \text{c}(22) * \text{Consonants} + \text{c}(24) * \text{Stress} \tag{9}$$

$$\text{Stress} = \text{c}(31) + \text{c}(33) * \text{Vowels} + \text{c}(34) * \text{Tonal} \tag{10}$$

where we omitted the coefficients that correspond to Stress in the consonant equation, Tonal in the vowel equation, Vowels in the tone equation, and Consonants in the stress equation.

Note that these results turned out to be considerably different than the ones that we obtained in 2014, in which we got a system where the stress equation had statistically significant coefficients for the other three variables, and the other equations only had significant coefficients for Stress. Besides, the author only obtained negative variable coefficients, while, in this new empirical exercise, four out of the eight estimated coefficients became positive[1].

One additional problem that these estimations have

Table 3. Main regression results for the original system of equations.

Concept	Regression 1		Regression 2		Regression 3	
	Coefficient	**Probab**	**Coefficient**	**Probab**	**Coefficient**	**Probab**
Consonant equation						
Vowels	1.28205	0.0000	1.25861	0.0000	0.89913	0.0000
Tone Distinction	5.59132	0.0078	4.68298	0.0125		
Stress Distinction	2.47514	0.2638				
R-squared	0.0477		0.0524		0.5581	
Vowel equation						
Consonants	0.17270	0.0000	0.17840	0.0000	0.18572	0.0000
Tone Distinction	0.93804	0.2259				
Stress Distinction	−1.82769	0.0228	−2.30636	0.0012		
R-squared	0.0687		0.0589		0.3372	
Tone equation						
Consonants	0.00818	0.0078	0.00891	0.0011		
Vowels	0.01019	0.2259				
Stress Distinction	−0.67962	0.0000	−0.69607	0.0000	−0.33447	0.0000
R-squared	0.1230		0.1169		0.5716	
Stress equation						
Consonants	0.00336	0.2638				
Vowels	−0.01840	0.0228	−0.01626	0.0239		
Tone Distinction	−0.62997	0.0000	−0.62193	0.0000	−0.52230	0.0000
R-squared	0.1070		0.1100		0.2646	

(and this is shared with the original results found in Coloma[1]) is that they have a relatively poor fit, in the sense that their R2 coefficients of determination are all closer to zero than to one. For an explanation of the use of R2 coefficients to interpret the goodness-of-fit of a regression, see Rasinger[10], chapter 7.

All this implies that there are probably many other factors that influence the values of the phonological variables, which are not included in our regressions. In order to improve that, the author decided to use an additional set of 29 categorical non-linguistic variables from the database, which relate to geographic, phylogenetic and demographic factors. All these variables are binary, and they indicate if a certain language belongs to a group that possesses a particular characteristic. The geographic variables correspond to the different macro-areas (North America, South America, East Africa, West Africa, Europe, West Asia, East Asia and Australasia), while the phylogenetic variables are related to the families represented by 3 or more languages in the sample (Niger-Congo, Indo-European, Austronesian, Sino-Tibetan, Afro-Asiatic, Altaic, Nilo-Saharan, Austro-Asiatic, Uralic, Trans-New Guinea, Khoisan, Dravidian, Tai-Kadai, Pama-Nyungan, Uto-Aztecan, Oto-Manguean and Arawakan). We also included two additional regional variables for the Caucasian languages (Kabardian, Georgian, Chechen, Lezgian and Armenian) and the Amazonian languages (Macushi, Yanomami, Guajajara, Xavante, Ticuna, Aguaruna and Shipibo), and two demographic variables for the Major languages and the Minor languages.

With that inclusion, the author ran a new system of equations whose fit improved considerably (Regression 3, with the main results reported in the last two columns of **Table 3**). However, this improvement had the undesired consequence that the only original coefficients that remained statistically significant were those relating consonants to vowels (and *vice versa*) and those relating tone to stress (and *vice versa*).

The final estimated system is therefore as follows:

$$\begin{aligned}\mathrm{Consonants} = {} & \mathrm{c}(1) * \mathrm{NorthAmerica} + (2) * \mathrm{SouthAmerica} + \mathrm{c}(3) * \mathrm{EastAfrica} + \mathrm{c}(4) * \mathrm{WestAfrica} \\ & + \mathrm{c}(5) * \mathrm{Europe} + \mathrm{c}(6) * \mathrm{WestAsia} + \mathrm{c}(7) * \mathrm{EastAsia} + \mathrm{c}(8) * \mathrm{Australasia} \\ & + \mathrm{c}(9) * \mathrm{NigerCongo} + \mathrm{c}(10) * \mathrm{IndoEuropean} + \mathrm{c}(11) * \mathrm{Austronesian}\end{aligned}$$

$$\begin{aligned}
&+\mathrm{c}(12) \ast \mathrm{SinoTibetan} + \mathrm{c}(13) \ast \mathrm{AfroAsiatic} + \mathrm{c}(14) \ast \mathrm{Altaic} + \mathrm{c}(15) \ast \mathrm{NiloSaharan}\\
&+\mathrm{c}(16) \ast \mathrm{AustroAsiatic} + \mathrm{c}(17) \ast \mathrm{Uralic} + \mathrm{c}(18) \ast \mathrm{TransNewGuinea} + \mathrm{c}(19) \ast \mathrm{Khoisan}\\
&+\mathrm{c}(20) \ast \mathrm{Dravidian} + \mathrm{c}(21) \ast \mathrm{TaiKadai} + \mathrm{c}(22) \ast \mathrm{PamaNyungan} + \mathrm{c}(23) \ast \mathrm{UtoAztecan}\\
&+\mathrm{c}(24) \ast \mathrm{OtoManguean} + \mathrm{c}(25) \ast \mathrm{Arawakan} + \mathrm{c}(26) \ast \mathrm{Caucasian} + \mathrm{c}(27) \ast \mathrm{Amazonian}\\
&+\mathrm{c}(28) \ast \mathrm{Major} + \mathrm{c}(29) \ast \mathrm{Minor} + \mathrm{c}(202) \ast \mathrm{Vowels}
\end{aligned} \tag{11}$$

$$\begin{aligned}
\mathrm{Vowels} = {}& \mathrm{c}(31) \ast \mathrm{NorthAmerica} + (32) \ast \mathrm{SouthAmerica} + \mathrm{c}(33) \ast \mathrm{EastAfrica} + \mathrm{c}(34) \ast \mathrm{WestAfrica}\\
&+\mathrm{c}(35) \ast \mathrm{Europe} + \mathrm{c}(36) \ast \mathrm{WestAsia} + \mathrm{c}(37) \ast \mathrm{EastAsia} + \mathrm{c}(38) \ast \mathrm{Australasia}\\
&+\mathrm{c}(39) \ast \mathrm{NigerCongo} + \mathrm{c}(40) \ast \mathrm{IndoEuropean} + \mathrm{c}(41) \ast \mathrm{Austronesian}\\
&+\mathrm{c}(42) \ast \mathrm{SinoTibetan} + \mathrm{c}(43) \ast \mathrm{AfroAsiatic} + \mathrm{c}(44) \ast \mathrm{Altaic} + \mathrm{c}(45) \ast \mathrm{NiloSaharan}\\
&+\mathrm{c}(46) \ast \mathrm{AustroAsiatic} + \mathrm{c}(47) \ast \mathrm{Uralic} + \mathrm{c}(48) \ast \mathrm{TransNewGuinea} + \mathrm{c}(49) \ast \mathrm{Khoisan}\\
&+\mathrm{c}(50) \ast \mathrm{Dravidian} + \mathrm{c}(51) \ast \mathrm{TaiKadai} + \mathrm{c}(52) \ast \mathrm{PamaNyungan} + \mathrm{c}(53) \ast \mathrm{UtoAztecan}\\
&+\mathrm{c}(54) \ast \mathrm{OtoManguean} + \mathrm{c}(55) \ast \mathrm{Arawakan} + \mathrm{c}(56) \ast \mathrm{Caucasian} + \mathrm{c}(57) \ast \mathrm{Amazonian}\\
&+\mathrm{c}(58) \ast \mathrm{Major} + \mathrm{c}(59) \ast \mathrm{Minor} + \mathrm{c}(212) \ast \mathrm{Consonants}
\end{aligned} \tag{12}$$

$$\begin{aligned}
\mathrm{Tonal} = {}& \mathrm{c}(61) \ast \mathrm{NorthAmerica} + (62) \ast \mathrm{SouthAmerica} + \mathrm{c}(63) \ast \mathrm{EastAfrica} + \mathrm{c}(64) \ast \mathrm{WestAfrica}\\
&+\mathrm{c}(65) \ast \mathrm{Europe} + \mathrm{c}(66) \ast \mathrm{WestAsia} + \mathrm{c}(67) \ast \mathrm{EastAsia} + \mathrm{c}(68) \ast \mathrm{Australasia}\\
&+\mathrm{c}(69) \ast \mathrm{NigerCongo} + \mathrm{c}(70) \ast \mathrm{IndoEuropean} + \mathrm{c}(71) \ast \mathrm{Austronesian}\\
&+\mathrm{c}(72) \ast \mathrm{SinoTibetan} + \mathrm{c}(73) \ast \mathrm{AfroAsiatic} + \mathrm{c}(74) \ast \mathrm{Altaic} + \mathrm{c}(75) \ast \mathrm{NiloSaharan}\\
&+\mathrm{c}(76) \ast \mathrm{AustroAsiatic} + \mathrm{c}(77) \ast \mathrm{Uralic} + \mathrm{c}(78) \ast \mathrm{TransNewGuinea} + \mathrm{c}(79) \ast \mathrm{Khoisan}\\
&+\mathrm{c}(80) \ast \mathrm{Dravidian} + \mathrm{c}(81) \ast \mathrm{TaiKadai} + \mathrm{c}(82) \ast \mathrm{PamaNyungan} + \mathrm{c}(83) \ast \mathrm{UtoAztecan}\\
&+\mathrm{c}(84) \ast \mathrm{OtoManguean} + \mathrm{c}(85) \ast \mathrm{Arawakan} + \mathrm{c}(86) \ast \mathrm{Caucasian} + \mathrm{c}(87) \ast \mathrm{Amazonian}\\
&+\mathrm{c}(88) \ast \mathrm{Major} + \mathrm{c}(89) \ast \mathrm{Minor} + \mathrm{c}(224) \ast \mathrm{Stress}
\end{aligned} \tag{13}$$

$$\begin{aligned}
\mathrm{Stress} = {}& \mathrm{c}(91) \ast \mathrm{NorthAmerica} + (92) \ast \mathrm{SouthAmerica} + \mathrm{c}(93) \ast \mathrm{EastAfrica} + \mathrm{c}(94) \ast \mathrm{WestAfrica}\\
&+\mathrm{c}(95) \ast \mathrm{Europe} + \mathrm{c}(996) \ast \mathrm{WestAsia} + \mathrm{c}(7) \ast \mathrm{EastAsia} + \mathrm{c}(98) \ast \mathrm{Australasia}\\
&+\mathrm{c}(99) \ast \mathrm{NigerCongo} + \mathrm{c}(100) \ast \mathrm{IndoEuropean} + \mathrm{c}(101) \ast \mathrm{Austronesian}\\
&+\mathrm{c}(102) \ast \mathrm{SinoTibetan} + \mathrm{c}(103) \ast \mathrm{AfroAsiatic} + \mathrm{c}(104) \ast \mathrm{Altaic} + \mathrm{c}(105) \ast \mathrm{NiloSaharan}\\
&+\mathrm{c}(106) \ast \mathrm{AustroAsiatic} + \mathrm{c}(107) \ast \mathrm{Uralic} + \mathrm{c}(108) \ast \mathrm{TransNewGuinea}\\
&+\mathrm{c}(109) \ast \mathrm{Khoisan} + \mathrm{c}(110) \ast \mathrm{Dravidian} + \mathrm{c}(111) \ast \mathrm{TaiKadai} + \mathrm{c}(112) \ast \mathrm{PamaNyungan}\\
&+\mathrm{c}(113) \ast \mathrm{UtoAztecan} + \mathrm{c}(114) \ast \mathrm{OtoManguean} + \mathrm{c}(115) \ast \mathrm{Arawakan}\\
&+\mathrm{c}(116) \ast \mathrm{Caucasian} + \mathrm{c}(117) \ast \mathrm{Amazonian} + \mathrm{c}(118) \ast \mathrm{Major} + \mathrm{c}(119) \ast \mathrm{Minor}\\
&+\mathrm{c}(234) \ast \mathrm{Tonal}
\end{aligned} \tag{14}$$

and the corresponding coefficients are positive for the relationships between vowels and consonants (c(202) and c(212)), and negative for the relationships between tone and stress (c(224) and c(234)).

4.2. Application to the New Model

As the database has information about the number of vowel qualities in each language (VowQual), together with information about the number of distinctive tones (Tones), and two categorical variables related to vowel nasalization (VowNasal) and vowel length (VowLength), the author has been able to estimate a new model with six equations, whose initial structure is as follows:

$$\mathrm{Consonants} = \mathrm{c}(1) + \mathrm{c}(2) \ast \mathrm{VowQual} + \mathrm{c}(3) \ast \mathrm{Tones} + \mathrm{c}(4) \ast \mathrm{Stress} + \mathrm{c}(5) \ast \mathrm{VowNasal} + \mathrm{c}(6) \ast \mathrm{VowLength} \tag{15}$$

$$\mathrm{VowQual} = \mathrm{c}(11) + \mathrm{c}(12) \ast \mathrm{Consonants} + \mathrm{c}(13) \ast \mathrm{Tones} + \mathrm{c}(14) \ast \mathrm{Stress} + \mathrm{c}(15) \ast \mathrm{VowNasal} + \mathrm{c}(16) \ast \mathrm{VowLength} \tag{16}$$

$$\mathrm{Tones} = \mathrm{c}(21) + \mathrm{c}(22) \ast \mathrm{Consonants} + \mathrm{c}(23) \ast \mathrm{VowQual} + \mathrm{c}(24) \ast \mathrm{Stress} + \mathrm{c}(25) \ast \mathrm{VowNasal} + \mathrm{c}(26) \ast \mathrm{VowLength} \tag{17}$$

$$\mathrm{Stress} = \mathrm{c}(31) + \mathrm{c}(32) \ast \mathrm{Consonants} + \mathrm{c}(33) \ast \mathrm{VowQual} + \mathrm{c}(34) \ast \mathrm{Tones} + \mathrm{c}(35) \ast \mathrm{VowNasal} + \mathrm{c}(36) \ast \mathrm{VowLength} \tag{18}$$

$$\text{VowNasal} = \text{c}(41) + \text{c}(42) \ast \text{Consonants} + \text{c}(43) \ast \text{VowQual} + \text{c}(44) \ast \text{Tones} + \text{c}(45) \ast \text{Stress} + \text{c}(46) \ast \text{VowLength} \tag{19}$$

$$\text{VowLength} = \text{c}(51) + \text{c}(52) \ast \text{Consonants} + \text{c}(53) \ast \text{VowQual} + \text{c}(54) \ast \text{Tones} + \text{c}(55) \ast \text{Stress} + \text{c}(56) \ast \text{VowNasal} \tag{20}$$

Following the same procedure used in the previous section (i.e., running SURs for the whole system), the author could identify a number of statistically significant coefficients, which defined a restricted system of equations as:

$$\text{Consonants} = \text{c}(1) + \text{c}(3) \ast \text{Tones} + \text{c}(5) \ast \text{VowNasal} \tag{21}$$

$$\text{VowQual} = \text{c}(11) + \text{c}(13) \ast \text{Tones} + \text{c}(14) \ast \text{Stress} + \text{c}(16) \ast \text{VowLength} \tag{22}$$

$$\text{Tones} = \text{c}(21) + \text{c}(22) \ast \text{Consonants} + \text{c}(23) \ast \text{VowQual} + \text{c}(24) \ast \text{Stress} \tag{23}$$

$$\text{Stress} = \text{c}(31) + \text{c}(33) \ast \text{VowQual} + \text{c}(34) \ast \text{Tones} + \text{c}(36) \ast \text{VowLength} \tag{24}$$

$$\text{VowNasal} = \text{c}(41) + \text{c}(42) \ast \text{Consonants} \tag{25}$$

$$\text{VowLength} = \text{c}(51) + \text{c}(53) \ast \text{VowQual} + \text{c}(55) \ast \text{Stress} \tag{26}$$

The main results for these restricted regression equations (Regression 4) are presented in the first three columns of **Table 4**. In it we see that all the estimated coefficients are statistically significant at a 5% probability level, but we also observe that the estimations have extremely low R^2 coefficients. That is why the author decided to include the same set of 29 non-linguistic categorical variables that were applied in the previous section, and, after eliminating the coefficients that were not statistically significant, this produced the results reported in the last three columns of **Table 4** (Regression 5).

The estimated version of the used system of regression equations with the additional categorical variables adopted the following form:

$$\begin{aligned}\text{Consonants} = {} & \text{c}(1) \ast \text{NorthAmerica} + (2) \ast \text{SouthAmerica} + \text{c}(3) \ast \text{EastAfrica} + \text{c}(4) \ast \text{WestAfrica} \\ & + \text{c}(5) \ast \text{Europe} + \text{c}(6) \ast \text{WestAsia} + \text{c}(7) \ast \text{EastAsia} + \text{c}(8) \ast \text{Australasia} \\ & + \text{c}(9) \ast \text{NigerCongo} + \text{c}(10) \ast \text{IndoEuropean} + \text{c}(11) \ast \text{Austronesian} \\ & + \text{c}(12) \ast \text{SinoTibetan} + \text{c}(13) \ast \text{AfroAsiatic} + \text{c}(14) \ast \text{Altaic} + \text{c}(15) \ast \text{NiloSaharan} \\ & + \text{c}(16) \ast \text{AustroAsiatic} + \text{c}(17) \ast \text{Uralic} + \text{c}(18) \ast \text{TransNewGuinea} + \text{c}(19) \ast \text{Khoisan} \\ & + \text{c}(20) \ast \text{Dravidian} + \text{c}(21) \ast \text{TaiKadai} + \text{c}(22) \ast \text{PamaNyungan} + \text{c}(23) \ast \text{UtoAztecan} \\ & + \text{c}(24) \ast \text{OtoManguean} + \text{c}(25) \ast \text{Arawakan} + \text{c}(26) \ast \text{Caucasian} + \text{c}(27) \ast \text{Amazonian} \\ & + \text{c}(28) \ast \text{Major} + \text{c}(29) \ast \text{Minor} + \text{c}(502) \ast \text{VowQual}\end{aligned} \tag{27}$$

$$\begin{aligned}\text{VowQual} = {} & \text{c}(31) \ast \text{NorthAmerica} + (32) \ast \text{SouthAmerica} + \text{c}(33) \ast \text{EastAfrica} + \text{c}(34) \ast \text{WestAfrica} \\ & + \text{c}(35) \ast \text{Europe} + \text{c}(36) \ast \text{WestAsia} + \text{c}(37) \ast \text{EastAsia} + \text{c}(38) \ast \text{Australasia} \\ & + \text{c}(39) \ast \text{NigerCongo} + \text{c}(40) \ast \text{IndoEuropean} + \text{c}(41) \ast \text{Austronesian} \\ & + \text{c}(42) \ast \text{SinoTibetan} + \text{c}(43) \ast \text{AfroAsiatic} + \text{c}(44) \ast \text{Altaic} + \text{c}(45) \ast \text{NiloSaharan} \\ & + \text{c}(46) \ast \text{AustroAsiatic} + \text{c}(47) \ast \text{Uralic} + \text{c}(48) \ast \text{TransNewGuinea} + \text{c}(49) \ast \text{Khoisan} \\ & + \text{c}(50) \ast \text{Dravidian} + \text{c}(51) \ast \text{TaiKadai} + \text{c}(52) \ast \text{PamaNyungan} + \text{c}(53) \ast \text{UtoAztecan} \\ & + \text{c}(54) \ast \text{OtoManguean} + \text{c}(55) \ast \text{Arawakan} + \text{c}(56) \ast \text{Caucasian} + \text{c}(57) \ast \text{Amazonian} \\ & + \text{c}(58) \ast \text{Major} + \text{c}(59) \ast \text{Minor} + \text{c}(512) \ast \text{Consonants} + \text{c}(515) \ast \text{VowNasal} \\ & + \text{c}(516) \ast \text{VowLength}\end{aligned} \tag{28}$$

$$\begin{aligned}\text{Tones} = {} & \text{c}(61) \ast \text{NorthAmerica} + (62) \ast \text{SouthAmerica} + \text{c}(63) \ast \text{EastAfrica} + \text{c}(64) \ast \text{WestAfrica} \\ & + \text{c}(65) \ast \text{Europe} + \text{c}(66) \ast \text{WestAsia} + \text{c}(67) \ast \text{EastAsia} + \text{c}(68) \ast \text{Australasia} \\ & + \text{c}(69) \ast \text{NigerCongo} + \text{c}(70) \ast \text{IndoEuropean} + \text{c}(71) \ast \text{Austronesian} \\ & + \text{c}(72) \ast \text{SinoTibetan} + \text{c}(73) \ast \text{AfroAsiatic} + \text{c}(74) \ast \text{Altaic} + \text{c}(75) \ast \text{NiloSaharan} \\ & + \text{c}(76) \ast \text{AustroAsiatic} + \text{c}(77) \ast \text{Uralic} + \text{c}(78) \ast \text{TransNewGuinea} + \text{c}(79) \ast \text{Khoisan} \\ & + \text{c}(80) \ast \text{Dravidian} + \text{c}(81) \ast \text{TaiKadai} + \text{c}(82) \ast \text{PamaNyungan} + \text{c}(83) \ast \text{UtoAztecan} \\ & + \text{c}(84) \ast \text{OtoManguean} + \text{c}(85) \ast \text{Arawakan} + \text{c}(86) \ast \text{Caucasian} + \text{c}(87) \ast \text{Amazonian}\end{aligned}$$

Table 4. Main regression results for the new systems of equations.

Concept		**Regression 4**			**Regression 5**	
	Coefficient	**t-Statistic**	**Probab**	**Coefficient**	**t-Statistic**	**Probab**
Consonant equation						
Vowel Qualities				1.49101	3.6431	0.0003
Tones	1.64212	2.8892	0.0040			
Vowel Nasalization	8.31007	3.5546	0.0004			
R-squared	0.0107			0.5585		
Vowel quality equation						
Consonants				0.05280	3.6976	0.0002
Tones	0.28287	2.8482	0.0045			
Stress Distinction	−1.04470	−2.7836	0.0055			
Vowel Nasalization				1.12534	3.1154	0.0019
Vowel Length	−2.14916	−6.6945	0.0000	−1.39436	−5.2095	0.0000
R-squared	0.0287			0.4864		
Tone equation						
Consonants	0.02461	2.4745	0.0135			
Vowel Qualities	0.17695	2.9966	0.0028			
Stress Distinction	−1.68632	−6.3829	0.0000	−1.23226	−4.9402	0.0000
R-squared	0.0648			0.4032		
Stress equation						
Vowel Qualities	−0.04750	−2.7854	0.0055			
Tones	−0.13525	−6.7432	0.0000	−0.12111	−5.1741	0.0000
Vowel Length	−0.39900	−5.9291	0.0000	−0.27849	−3.9496	0.0001
R-squared	0.0611			0.2741		
Nasalization equation						
Consonants	0.00983	3.6827	0.0002			
Vowel Qualities				0.05608	3.4381	0.0006
R-squared	0.0037			0.3833		
Length equation						
Vowel Qualities	−0.11883	−6.7857	0.0000	−0.11027	−5.2720	0.0000
Stress Distinction	−0.46449	−5.7396	0.0000	−0.30280	−3.7386	0.0002
R-squared	0.0290			0.2890		

$$+\mathrm{c}(88)*\mathrm{Major}+\mathrm{c}(89)*\mathrm{Minor}+\mathrm{c}(524)*\mathrm{Stress} \tag{29}$$

$$\begin{aligned}
\mathrm{Stress} = {} & \mathrm{c}(91)*\mathrm{NorthAmerica}+(92)*\mathrm{SouthAmerica}+\mathrm{c}(93)*\mathrm{EastAfrica}+\mathrm{c}(94)*\mathrm{WestAfrica}\\
& +\mathrm{c}(95)*\mathrm{Europe}+\mathrm{c}(996)*\mathrm{WestAsia}+\mathrm{c}(7)*\mathrm{EastAsia}+\mathrm{c}(98)*\mathrm{Australasia}\\
& +\mathrm{c}(99)*\mathrm{NigerCongo}+\mathrm{c}(100)*\mathrm{IndoEuropean}+\mathrm{c}(101)*\mathrm{Austronesian}\\
& +\mathrm{c}(102)*\mathrm{SinoTibetan}+\mathrm{c}(103)*\mathrm{AfroAsiatic}+\mathrm{c}(104)*\mathrm{Altaic}+\mathrm{c}(105)*\mathrm{NiloSaharan}\\
& +\mathrm{c}(106)*\mathrm{AustroAsiatic}+\mathrm{c}(107)*\mathrm{Uralic}+\mathrm{c}(108)*\mathrm{TransNewGuinea}\\
& +\mathrm{c}(109)*\mathrm{Khoisan}+\mathrm{c}(110)*\mathrm{Dravidian}+\mathrm{c}(111)*\mathrm{TaiKadai}+\mathrm{c}(112)*\mathrm{PamaNyungan}\\
& +\mathrm{c}(113)*\mathrm{UtoAztecan}+\mathrm{c}(114)*\mathrm{OtoManguean}+\mathrm{c}(115)*\mathrm{Arawakan}\\
& +\mathrm{c}(116)*\mathrm{Caucasian}+\mathrm{c}(117)*\mathrm{Amazonian}+\mathrm{c}(118)*\mathrm{Major}+\mathrm{c}(119)*\mathrm{Minor}\\
& +\mathrm{c}(534)*\mathrm{Tones}+\mathrm{c}(536)*\mathrm{VowLength}
\end{aligned} \tag{30}$$

$$\begin{aligned}
\mathrm{VowNasal} = {} & \mathrm{c}(121)*\mathrm{NorthAmerica}+(122)*\mathrm{SouthAmerica}+\mathrm{c}(123)*\mathrm{EastAfrica}\\
& +\mathrm{c}(124)*\mathrm{WestAfrica}+\mathrm{c}(125)*\mathrm{Europe}+\mathrm{c}(126)*\mathrm{WestAsia}+\mathrm{c}(127)*\mathrm{EastAsia}\\
& +\mathrm{c}(128)*\mathrm{Australasia}+\mathrm{c}(129)*\mathrm{NigerCongo}+\mathrm{c}(130)*\mathrm{IndoEuropean}
\end{aligned}$$

$$\begin{aligned}
&+c(131)\ast\text{Austronesian}+c(132)\ast\text{SinoTibetan}+c(133)\ast\text{AfroAsiatic}+c(134)\ast\text{Altaic}\\
&+c(135)\ast\text{NiloSaharan}+c(136)\ast\text{AustroAsiatic}+c(137)\ast\text{Uralic}\\
&+c(138)\ast\text{TransNewGuinea}+c(139)\ast\text{Khoisan}+c(140)\ast\text{Dravidian}\\
&+c(141)\ast\text{TaiKadai}+c(142)\ast\text{PamaNyungan}+c(143)\ast\text{UtoAztecan}\\
&+c(144)\ast\text{OtoManguean}+c(145)\ast\text{Arawakan}+c(146)\ast\text{Caucasian}\\
&+c(147)\ast\text{Amazonian}+c(148)\ast\text{Major}+c(149)\ast\text{Minor}+c(543)\ast\text{VowQual}
\end{aligned}\tag{31}$$

$$\begin{aligned}
\text{VowLength}=\ &c(151)\ast\text{NorthAmerica}+(152)\ast\text{SouthAmerica}+c(153)\ast\text{EastAfrica}\\
&+c(154)\ast\text{WestAfrica}+c(155)\ast\text{Europe}+c(156)\ast\text{WestAsia}+c(157)\ast\text{EastAsia}\\
&+c(158)\ast\text{Australasia}+c(159)\ast\text{NigerCongo}+c(160)\ast\text{IndoEuropean}\\
&+c(161)\ast\text{Austronesian}+c(162)\ast\text{SinoTibetan}+c(163)\ast\text{AfroAsiatic}+c(164)\ast\text{Altaic}\\
&+c(165)\ast\text{NiloSaharan}+c(166)\ast\text{AustroAsiatic}+c(167)\ast\text{Uralic}\\
&+c(168)\ast\text{TransNewGuinea}+c(169)\ast\text{Khoisan}+c(170)\ast\text{Dravidian}\\
&+c(171)\ast\text{TaiKadai}+c(172)\ast\text{PamaNyungan}+c(173)\ast\text{UtoAztecan}\\
&+c(174)\ast\text{OtoManguean}+c(175)\ast\text{Arawakan}+c(176)\ast\text{Caucasian}\\
&+c(177)\ast\text{Amazonian}+c(178)\ast\text{Major}+c(179)\ast\text{Minor}+c(553)\ast\text{VowQual}\\
&+c(555)\ast\text{Stress}
\end{aligned}\tag{32}$$

and all the estimated coefficients for the phonological variables (c(502), c(512), c(515), c(516), c(524), c(534), c(536), c(543), c(553) and c(555)) became statistically significant at a 1% probability level, while the corresponding R2 coefficients improved considerably (compared to the ones calculated for Regression 4).

As observed in the last three columns of **Table 4**, the signs of the regression coefficients indicate a positive correlation for Consonants/VowQual and VowQual/VowNasal, and a negative correlation for VowQual/VowLength, Tones/Stress, and Stress/VowLength.

4.3. Welfare Function Estimation

The system formed by Equations (27) to (32) can be used to estimate a welfare function for our sample of languages. This is done by rewriting that system as follows:

$$\begin{aligned}
\text{Consonants}=\ &c(201)\ast(\text{NorthAmerica}+(2)\ast\text{SouthAmerica}+c(3)\ast\text{EastAfrica}+c(4)\ast\text{WestAfrica}\\
&+c(5)\ast\text{Europe}+c(6)\ast\text{WestAsia}+c(7)\ast\text{EastAsia}+c(8)\ast\text{Australasia}\\
&+c(9)\ast\text{NigerCongo}+c(10)\ast\text{IndoEuropean}+c(11)\ast\text{Austronesian}\\
&+c(12)\ast\text{SinoTibetan}+c(13)\ast\text{AfroAsiatic}+c(14)\ast\text{Altaic}+c(15)\ast\text{NiloSaharan}\\
&+c(16)\ast\text{AustroAsiatic}+c(17)\ast\text{Uralic}+c(18)\ast\text{TransNewGuinea}+c(19)\ast\text{Khoisan}\\
&+c(20)\ast\text{Dravidian}+c(21)\ast\text{TaiKadai}+c(22)\ast\text{PamaNyungan}+c(23)\ast\text{UtoAztecan}\\
&+c(24)\ast\text{OtoManguean}+c(25)\ast\text{Arawakan}+c(26)\ast\text{Caucasian}+c(27)\ast\text{Amazonian}\\
&+c(28)\ast\text{Major}+c(29)\ast\text{Minor})+c(201)\ast c(202)\ast\text{VowQual}
\end{aligned}\tag{33}$$

$$\begin{aligned}
\text{VowQual}=\ &c(31)\ast\text{NorthAmerica}+(32)\ast\text{SouthAmerica}+c(33)\ast\text{EastAfrica}+c(34)\ast\text{WestAfrica}\\
&+c(35)\ast\text{Europe}+c(36)\ast\text{WestAsia}+c(37)\ast\text{EastAsia}+c(38)\ast\text{Australasia}\\
&+c(39)\ast\text{NigerCongo}+c(40)\ast\text{IndoEuropean}+c(41)\ast\text{Austronesian}\\
&+c(42)\ast\text{SinoTibetan}+c(43)\ast\text{AfroAsiatic}+c(44)\ast\text{Altaic}+c(45)\ast\text{NiloSaharan}\\
&+c(46)\ast\text{AustroAsiatic}+c(47)\ast\text{Uralic}+c(48)\ast\text{TransNewGuinea}+c(49)\ast\text{Khoisan}\\
&+c(50)\ast\text{Dravidian}+c(51)\ast\text{TaiKadai}+c(52)\ast\text{PamaNyungan}+c(53)\ast\text{UtoAztecan}\\
&+c(54)\ast\text{OtoManguean}+c(55)\ast\text{Arawakan}+c(56)\ast\text{Caucasian}+c(57)\ast\text{Amazonian}\\
&+c(58)\ast\text{Major}+c(59)\ast\text{Minor}+c(212)\ast c(202)\ast\text{Consonants}\\
&+c(212)\ast c(215)\ast\text{VowNasal}+c(212)\ast c(216)\ast\text{VowLength}
\end{aligned}\tag{34}$$

$$\text{Tones}=c(61)\ast\text{NorthAmerica}+(62)\ast\text{SouthAmerica}+c(63)\ast\text{EastAfrica}+c(64)\ast\text{WestAfrica}$$

$$\begin{aligned}
&+\mathrm{c}(65)*\mathrm{Europe}+\mathrm{c}(66)*\mathrm{WestAsia}+\mathrm{c}(67)*\mathrm{EastAsia}+\mathrm{c}(68)*\mathrm{Australasia}\\
&+\mathrm{c}(69)*\mathrm{NigerCongo}+\mathrm{c}(70)*\mathrm{IndoEuropean}+\mathrm{c}(71)*\mathrm{Austronesian}\\
&+\mathrm{c}(72)*\mathrm{SinoTibetan}+\mathrm{c}(73)*\mathrm{AfroAsiatic}+\mathrm{c}(74)*\mathrm{Altaic}+\mathrm{c}(75)*\mathrm{NiloSaharan}\\
&+\mathrm{c}(76)*\mathrm{AustroAsiatic}+\mathrm{c}(77)*\mathrm{Uralic}+\mathrm{c}(78)*\mathrm{TransNewGuinea}+\mathrm{c}(79)*\mathrm{Khoisan}\\
&+\mathrm{c}(80)*\mathrm{Dravidian}+\mathrm{c}(81)*\mathrm{TaiKadai}+\mathrm{c}(82)*\mathrm{PamaNyungan}+\mathrm{c}(83)*\mathrm{UtoAztecan}\\
&+\mathrm{c}(84)*\mathrm{OtoManguean}+\mathrm{c}(85)*\mathrm{Arawakan}+\mathrm{c}(86)*\mathrm{Caucasian}+\mathrm{c}(87)*\mathrm{Amazonian}\\
&+\mathrm{c}(88)*\mathrm{Major}+\mathrm{c}(89)*\mathrm{Minor}+\mathrm{c}(222)*\mathrm{c}(224)*\mathrm{Stress}
\end{aligned}\tag{35}$$

$$\begin{aligned}
\mathrm{Stress}=\;&\mathrm{c}(91)*\mathrm{NorthAmerica}+(92)*\mathrm{SouthAmerica}+\mathrm{c}(93)*\mathrm{EastAfrica}+\mathrm{c}(94)*\mathrm{WestAfrica}\\
&+\mathrm{c}(95)*\mathrm{Europe}+\mathrm{c}(996)*\mathrm{WestAsia}+\mathrm{c}(7)*\mathrm{EastAsia}+\mathrm{c}(98)*\mathrm{Australasia}\\
&+\mathrm{c}(99)*\mathrm{NigerCongo}+\mathrm{c}(100)*\mathrm{IndoEuropean}+\mathrm{c}(101)*\mathrm{Austronesian}\\
&+\mathrm{c}(102)*\mathrm{SinoTibetan}+\mathrm{c}(103)*\mathrm{AfroAsiatic}+\mathrm{c}(104)*\mathrm{Altaic}+\mathrm{c}(105)*\mathrm{NiloSaharan}\\
&+\mathrm{c}(106)*\mathrm{AustroAsiatic}+\mathrm{c}(107)*\mathrm{Uralic}+\mathrm{c}(108)*\mathrm{TransNewGuinea}\\
&+\mathrm{c}(109)*\mathrm{Khoisan}+\mathrm{c}(110)*\mathrm{Dravidian}+\mathrm{c}(111)*\mathrm{TaiKadai}+\mathrm{c}(112)*\mathrm{PamaNyungan}\\
&+\mathrm{c}(113)*\mathrm{UtoAztecan}+\mathrm{c}(114)*\mathrm{OtoManguean}+\mathrm{c}(115)*\mathrm{Arawakan}\\
&+\mathrm{c}(116)*\mathrm{Caucasian}+\mathrm{c}(117)*\mathrm{Amazonian}+\mathrm{c}(118)*\mathrm{Major}+\mathrm{c}(119)*\mathrm{Minor}\\
&+\mathrm{c}(232)*\mathrm{c}(224)*\mathrm{Tones}+\mathrm{c}(232)*\mathrm{c}(236)*\mathrm{VowLength}
\end{aligned}\tag{36}$$

$$\begin{aligned}
\mathrm{VowNasal}=\;&\mathrm{c}(121)*\mathrm{NorthAmerica}+(122)*\mathrm{SouthAmerica}+\mathrm{c}(123)*\mathrm{EastAfrica}\\
&+\mathrm{c}(124)*\mathrm{WestAfrica}+\mathrm{c}(125)*\mathrm{Europe}+\mathrm{c}(126)*\mathrm{WestAsia}+\mathrm{c}(127)*\mathrm{EastAsia}\\
&+\mathrm{c}(128)*\mathrm{Australasia}+\mathrm{c}(129)*\mathrm{NigerCongo}+\mathrm{c}(130)*\mathrm{IndoEuropean}\\
&+\mathrm{c}(131)*\mathrm{Austronesian}+\mathrm{c}(132)*\mathrm{SinoTibetan}+\mathrm{c}(133)*\mathrm{AfroAsiatic}+\mathrm{c}(134)*\mathrm{Altaic}\\
&+\mathrm{c}(135)*\mathrm{NiloSaharan}+\mathrm{c}(136)*\mathrm{AustroAsiatic}+\mathrm{c}(137)*\mathrm{Uralic}\\
&+\mathrm{c}(138)*\mathrm{TransNewGuinea}+\mathrm{c}(139)*\mathrm{Khoisan}+\mathrm{c}(140)*\mathrm{Dravidian}\\
&+\mathrm{c}(141)*\mathrm{TaiKadai}+\mathrm{c}(142)*\mathrm{PamaNyungan}+\mathrm{c}(143)*\mathrm{UtoAztecan}\\
&+\mathrm{c}(144)*\mathrm{OtoManguean}+\mathrm{c}(145)*\mathrm{Arawakan}+\mathrm{c}(146)*\mathrm{Caucasian}\\
&+\mathrm{c}(147)*\mathrm{Amazonian}+\mathrm{c}(148)*\mathrm{Major}+\mathrm{c}(149)*\mathrm{Minor}\\
&+\mathrm{c}(242)*\mathrm{c}(215)*\mathrm{VowQual}
\end{aligned}\tag{37}$$

$$\begin{aligned}
\mathrm{VowLength}=\;&\mathrm{c}(151)*\mathrm{NorthAmerica}+(152)*\mathrm{SouthAmerica}+\mathrm{c}(153)*\mathrm{EastAfrica}\\
&+\mathrm{c}(154)*\mathrm{WestAfrica}+\mathrm{c}(155)*\mathrm{Europe}+\mathrm{c}(156)*\mathrm{WestAsia}+\mathrm{c}(157)*\mathrm{EastAsia}\\
&+\mathrm{c}(158)*\mathrm{Australasia}+\mathrm{c}(159)*\mathrm{NigerCongo}+\mathrm{c}(160)*\mathrm{IndoEuropean}\\
&+\mathrm{c}(161)*\mathrm{Austronesian}+\mathrm{c}(162)*\mathrm{SinoTibetan}+\mathrm{c}(163)*\mathrm{AfroAsiatic}+\mathrm{c}(164)*\mathrm{Altaic}\\
&+\mathrm{c}(165)*\mathrm{NiloSaharan}+\mathrm{c}(166)*\mathrm{AustroAsiatic}+\mathrm{c}(167)*\mathrm{Uralic}\\
&+\mathrm{c}(168)*\mathrm{TransNewGuinea}+\mathrm{c}(169)*\mathrm{Khoisan}+\mathrm{c}(170)*\mathrm{Dravidian}\\
&+\mathrm{c}(171)*\mathrm{TaiKadai}+\mathrm{c}(172)*\mathrm{PamaNyungan}+\mathrm{c}(173)*\mathrm{UtoAztecan}\\
&+\mathrm{c}(174)*\mathrm{OtoManguean}+\mathrm{c}(175)*\mathrm{Arawakan}+\mathrm{c}(176)*\mathrm{Caucasian}\\
&+\mathrm{c}(177)*\mathrm{Amazonian}+\mathrm{c}(178)*\mathrm{Major}+\mathrm{c}(179)*\mathrm{Minor}\\
&+\mathrm{c}(252)*\mathrm{c}(216)*\mathrm{VowQual}+\mathrm{c}(252)*\mathrm{c}(236)*\mathrm{Stress}
\end{aligned}\tag{38}$$

This new system, formed by Equations (33) to (38), includes a number of relationships between the parameters of the welfare function. These are given by coefficients c(202), c(215), c(216), c(224) and c(236), each of which appears in two different equations. These are precisely the coefficients that measure the a_{ij} parameters of the welfare function, which show the link between the different phonological variables in that function.

Running the regressions for Equations (33) to (38), using the same estimation procedure applied in the previous sections, yielded a new set of coefficients. This allowed the calculation of values of the implicit welfare function parameters, according to

a_1 = 1;
a_2 = Average(c(31) to c(59))/c(212) = 12.2871;
a_3 = Average(c(61) to c(89))/c(222) = 6.9535;
a_4 = Average(c(91) to c(119))/c(232) = 20.1793;
a_5 = Average(c(121) to c(149))/c(242) = –5.6752;
a_6 = Average(c(151) to c(179))/c(252) = 33.7650;
a_{12} = c(202) = 0.1308; a_{25} = c(215) = 2.4418; a_{26} = c(216) = –3.1043;
a_{34} = c(224) = –3.8530; a_{46} = c(236) = –9.2300; b_1 = 1/c(201) = 0.0750;
b_2 = 1/c(212) = 2.3968; b_3 = 1/c(222) = 3.1270; b_4 = 1/c(232) = 31.8452;
b_5 = 1/c(242) = 46.3680; b_6 = 1/c(252) = 30.0420.

The obtained results also indicate that parameters a_{13}, a_{14}, a_{15}, a_{16}, a_{23}, a_{24}, a_{35}, a_{36}, a_{45} and a_{56} are all equal to zero. Therefore, the estimated welfare function turns out to be

$$\begin{aligned} W = {} & \text{Consonants} + 12.2871 \ast \text{VowQual} + 6.9535 \ast \text{Tones} + 20.1793 \ast \text{Stress} – 5.6752 \ast \text{VowNasal} \\ & + 33.765 \ast \text{VowLength} + 0.1308 \ast \text{Consonants} \ast \text{VowQual} \\ & + 2.4418 \ast \text{VowQual} \ast \text{VowNasal} – 3.1043 \ast \text{VowQual} \ast \text{VowLength} \\ & – 3.853 \ast \text{Tones} \ast \text{Stress} – 9.23 \ast \text{Stress} \ast \text{VowLength} \\ & – 0.5 \ast (0.075 \ast \text{Consonants2} + 2.3968 \ast \text{VowQual2} + 3.127 \ast \text{Tones2} \\ & + 31.8452 \ast \text{Stress2} + 46.368 \ast \text{VowNasal2} + 30.042 \ast \text{VowLength2}) \end{aligned} \tag{39}$$

Note that, by definition, it is assumed that "a_1 = 1", because the parameter that corresponds to the consonants' decoding ease as a *numéraire* for the whole system is used. This is due to the fact that welfare is a concept whose measure is arbitrary, so any linear transformation of the welfare function is fine to be used in this context.

The parameters of the welfare function can also generate a set of "partial correlation coefficients" for the whole system. These coefficients are equal to zero for the pairs of phonological variables for which it holds that "a_{ij} = 0" but, for the five cases that display non-zero parameters, the corresponding correlation coefficients (r_{ij}) can be calculated using

$$r_{ij} = \pm\sqrt{\frac{a_{ij}^2}{b_i \cdot b_j}} \tag{40}$$

where the sign of r_{ij} is positive or negative, depending on the sign of each a_{ij} parameter.

The newly calculated correlation coefficients are reported in **Table 5**. When comparing them with the standard correlation coefficients in **Table 2**, it is found that the absolute values of the new coefficients are all higher than those of the original ones. This is because these figures are based on an estimation that was limited to capturing only the most significant relationships among the phonological variables of our system.

Another way to represent the different parts of the estimated welfare function is the one presented in **Figure 6**. **Figure 6** depicts the values of that function for varying numbers of consonant phonemes, assuming that all the other variables took a value equal to their average in the whole sample. As observed, the difference between decoding ease and production effort is maximal when a language has 24.39 consonants, which is precisely the average number of consonant phonemes in our database. This is because all the parameters of the welfare function were calibrated by the regression procedure, in order to represent a situation in which welfare is maximized at the actual average values of the phonological variables included in the function.

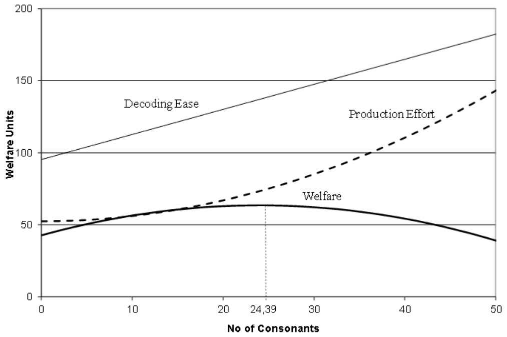

Figure 6. Decoding ease, production effort and welfare.

Table 5. Partial correlation coefficients from the welfare function.

Concept	Consonants	VowQual	Tones	Stress	VowNasal	VowLength
Consonants	1.0000					
Vowel Qualities	0.2968	1.0000				
Tones			1.0000			
Stress			−0.3863	1.0000		
Vowel Nasalization		0.2512			1.0000	
Vowel Length		−0.3921		−0.2904		1.0000

The interaction between decoding ease and production effort in the maximization of welfare can also be seen in **Figure 7**, where both concepts have been represented in the space of consonant phonemes versus vowel qualities. As we see, the point where "Consonants = 24.39" and "Vowel Qualities = 6.11" (i.e., the average values for the whole sample) is the one where decoding ease is maximal (D = 137.75) for a certain level of production effort (P = 74.08), and it is also the place where production effort is minimal for that given level of decoding ease.

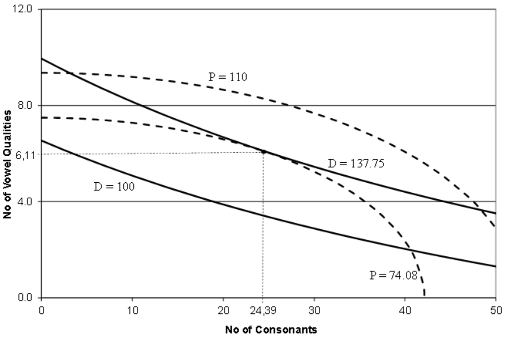

Figure 7. Decoding ease and production effort for different consonants and vowel qualities.

Of course, it would be possible to obtain the same decoding ease value with other combinations of consonant phonemes and vowel qualities (e.g., with three consonants and 9.36 vowel qualities, or with 3.67 vowel qualities and 48 consonants), but that would imply a larger production effort (P = 110). Similarly, it would be possible to choose other combinations where the level of production effort is "P = 74.08", but that would imply a lower level of decoding ease. Note that those alternative levels of decoding ease and production effort cross the other curves at relatively extreme positions in the depicted space (i.e., at points where there are many consonants and few vowel qualities, or *vice versa*). Conversely, in the point where welfare is maximized, the values of the phonological variables are more moderate, and the decoding ease line is tangent to the corresponding production effort curve.

In **Figure 8**, the same space of consonant phonemes and vowel qualities is used to represent different levels of the welfare function (i.e., of the difference between decoding ease and production effort). It can be seen that there are "indifference curves" or "iso-welfare curves", which represent points where the welfare function has the same value (which in this case is "W = 50", "W = 56", "W = 60" or "W = 63.67"). This last number is the value of the welfare function for the average values of the depicted phonological variables (i.e., "Consonants = 24.39" and "Vowel Qualities = 6.11"). What we observe is that, as we move away from those average values, welfare decreases, producing circles or ovals around its maximum value.

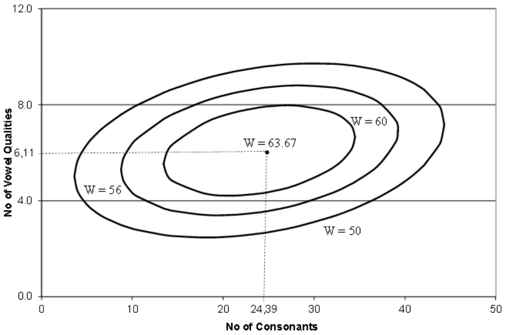

Figure 8. Welfare values for different consonants and vowel qualities.

5. Concluding Remarks

The primary conclusion drawn from the various analyses conducted in this paper is that the synergetic phonology model developed in 2014 is unable to match the new database, which comprises 150 languages. This database features a greater number of languages, language families, and phylogenetic and geographic diversity compared to the original 100-language database[1].

That original model, however, can be adapted by including some additional phonological variables (mainly, a variable that describes the tone structure in a more precise way, and three variables that describe the structure of the languages' vowel inventory with more detail). Some non-linguistic variables can also be included, related to the origins of the languages and their geographic and demographic characteristics. With those inclusions, a new empirical model can be derived, which can be represented by a diagram as shown in **Figure 9**.

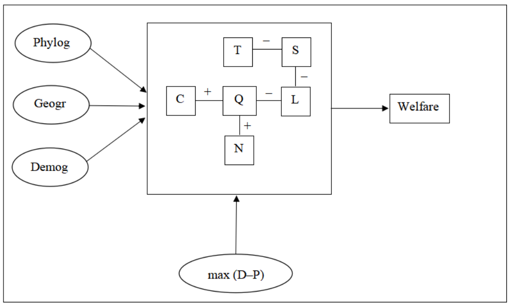

Figure 9. Diagram of the new estimated model.

Figure 9 is basically the same one presented in section 2, with the addition of certain explicit relationships detected between the phonological variables. It can be seen that the key variable now appears to be the number of vowel qualities (Q), which is related to the number of consonants (C), the presence of vowel nasalization (N), and the use of contrasting vowel length (L). That last variable is, in turn, linked to the existence of distinctive stress (S), which is also related to the number of distinctive tones (T).

The only relationship that remains from the original model is precisely that last link between stress and tone, which also keeps its negative and significant correlation in our new database. The other two relationships (between stress and vowels, and between stress and consonants) have now been replaced by other relationships, such as those mentioned in the previous paragraph.

Note that this new model allows for three negative correlations between phonological variables (and therefore explains some "trade-offs" between those variables), but it also finds positive correlation coefficients between consonants and vowel qualities, and between vowel qualities and vowel nasalization. This seems to signal the existence of some kind of complementarity between those variables, which is explainable through the idea that such variables, taken together, may contribute to increasing the decoding ease of language. The cases of negative correlation, conversely, can be interpreted as situations in which there is substitution between the phonological variables (e.g., the idea that, for the same purpose, some languages may use stress distinctions while others may use tone or vowel length distinctions).

In 2014, the author suggested the use of additional characteristics to expand the set of phonological variables as a possible avenue for future research. This is precisely what was done in this paper, along with the inclusion of non-linguistic variables to account for a series of exogenous factors. All this helped uncover new relationships between the variables, in a search guided by the use of a synergetic linguistics approach, based on the idea that languages try to maximize decoding ease and to minimize production effort.

Other possible lines of research mentioned involved the inclusion of morphological and syntactic variables[1]. That is something the author explored in other contributions[11–14], using different sets of typological variables, or empirical variables calculated from actual texts. Those sets were either based on the World Atlas of Language Structures[7] or on the text of the fable "The North Wind and the Sun", translated into different languages. This paper returned to the analysis of phonological variables only, but it made a much more careful selection of the included observations in order to have a more representative and complete language sample.

Funding

This work received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data used in this study is reproduced in **Appendix A** and **Appendix B**.

Acknowledgments

The author thanks Bahía Solla for her assistance in preparing this manuscript, and one anonymous reviewer for his/her useful comments.

Conflicts of Interest

The author declares no conflict of interest.

Appendix A

Table A1. Language list.

Language	Genus	Family	Area	Country/Region	Size
Aguaruna	Chicham	Jivaroan	South America	Peru	Minor
Akan	Kwa	Niger-Congo	West Africa	Ghana	Medium
Albanian	Albanic	Indo-European	Europe	Albania	Medium
Arabic	Semitic	Afro-Asiatic	West Asia	Saudi Arabia	Major
Armenian	Armenic	Indo-European	West Asia	Armenia	Medium
Arrernte	Arandic	Pama-Nyungan	Australasia	Australia	Minor
Asheninka	Pre-Andine	Arawakan	South America	Peru	Minor
Atayal	North Formosan	Austronesian	East Asia	Taiwan	Minor
Aymara	Aymaran	Andean	South America	Bolivia	Medium
Bambara	Mande	Niger-Congo	West Africa	Mali	Medium
Basque	Vasconic	Vasconic	Europe	Spain	Medium
Batak	Sumatran	Austronesian	Australasia	Indonesia	Medium
Beja	North Cushitic	Afro-Asiatic	East Africa	Sudan	Medium
Berber	Berberic	Afro-Asiatic	West Africa	Morocco	Medium
Blackfoot	West Algonquian	Algic	North America	Canada	Minor
Brahui	North Dravidian	Dravidian	West Asia	Pakistan	Medium
Bugis	South Sulawesi	Austronesian	Australasia	Indonesia	Medium
Burmese	Burmic	Sino-Tibetan	East Asia	Myanmar	Major
Burushaski	Burushaskian	Burushaskian	West Asia	Pakistan	Medium
Cambodian	Khmer	Austro-Asiatic	East Asia	Cambodia	Major
Chamorro	Chamorro	Austronesian	Australasia	Guam	Minor
Chechen	Nakh	East Caucasian	West Asia	Russia	Medium
Cherokee	South Iroquoian	Iroquoian	North America	United States	Minor
Choctaw	West Muskogean	Muskogean	North America	United States	Minor
Chukchi	Chukotkan	Paleo-Siberian	East Asia	Russia	Minor
Cree	Central Algonquian	Algic	North America	Canada	Minor
Dani	Irian Highland	Trans-New Guinea	Australasia	Indonesia	Medium
Dholuo	Nilotic	Nilo-Saharan	East Africa	Kenya	Medium
Dogon	Dogonic	Niger-Congo	West Africa	Mali	Medium
Embera	Chocoan	Chocoan	South America	Colombia	Minor
Enga	Engan	Trans-New Guinea	Australasia	Papua New Guinea	Medium
English	Germanic	Indo-European	Europe	United Kingdom	Major
Evenki	Tungusic	Altaic	East Asia	Russia	Minor
Fijian	Oceanic	Austronesian	Australasia	Fiji	Medium
Filipino	Central Philippine	Austronesian	Australasia	Philippines	Major
Finnish	Finnic	Uralic	Europe	Finland	Medium
Fulfulde	Senegambian	Niger-Congo	West Africa	Mali	Major
Fur	Fur	Nilo-Saharan	East Africa	Sudan	Medium
Garifuna	Caribbean	Arawakan	North America	Belize	Medium
Georgian	Kartvelian	South Caucasian	West Asia	Georgia	Medium
Greek	Hellenic	Indo-European	Europe	Greece	Major
Guajajara	Teneteharan	Tupian	South America	Brazil	Minor
Guarani	Guaranitic	Tupian	South America	Paraguay	Medium
Guaymi	Guaymiic	Chibchan	North America	Panama	Medium
Gumuz	Komuz	Nilo-Saharan	East Africa	Ethiopia	Medium

Table A1. *Cont.*

Language	Genus	Family	Area	Country/Region	Size
Gunwinggu	Gunwinggic	Gunwinyguan	Australasia	Australia	Minor
Haitian	French-based	Creole	North America	Haiti	Major
Hausa	West Chadic	Afro-Asiatic	West Africa	Nigeria	Major
Hindi	Indic	Indo-European	West Asia	India	Major
Hlai	Hlaic	Tai-Kadai	East Asia	China	Medium
Hmong	Hmongic	Hmong-Mien	East Asia	China	Medium
Hungarian	Ugric	Uralic	Europe	Hungary	Major
Iatmul	Sepik	North Papuan	Australasia	Papua New Guinea	Minor
Ibibio	Delta Cross	Niger-Congo	West Africa	Nigeria	Medium
Ijo	Ijoid	Niger-Congo	West Africa	Nigeria	Medium
Indonesian	Malayic	Austronesian	Australasia	Indonesia	Major
Inuit	Eskimoan	Eskimo-Aleut	North America	Canada	Minor
Iraqw	South Cushitic	Afro-Asiatic	East Africa	Tanzania	Medium
Irish	Celtic	Indo-European	Europe	Ireland	Medium
Japanese	Japonic	Altaic	East Asia	Japan	Major
Jinghpo	Kachin	Sino-Tibetan	East Asia	Myanmar	Medium
Kabardian	Circassian	West Caucasian	West Asia	Russia	Medium
Kabiye	Gur	Niger-Congo	West Africa	Togo	Medium
Kam	Kam-Sui	Tai-Kadai	East Asia	China	Medium
Kamano	Gorokan	Trans-New Guinea	Australasia	Papua New Guinea	Minor
Kanuri	Saharan	Nilo-Saharan	West Africa	Nigeria	Medium
Karen	Karenic	Sino-Tibetan	East Asia	Thailand	Medium
Kazakh	Kipchak Turkic	Altaic	West Asia	Kazakhstan	Major
Khoekhoe	Khoe-Kwadi	Khoisan	East Africa	Namibia	Medium
Kiche	Quichean	Mayan	North America	Guatemala	Medium
Korean	Koreanic	Altaic	East Asia	Korea	Major
Kuman	Chimbu-Wahgi	Trans-New Guinea	Australasia	Papua New Guinea	Medium
Kunama	Kunaman	Nilo-Saharan	East Africa	Eritrea	Medium
Lezgian	Lezgic	East Caucasian	West Asia	Russia	Medium
Lithuanian	Baltic	Indo-European	Europe	Lithuania	Medium
Macushi	Pemongan	Cariban	South America	Guyana	Minor
Madi	Central Sudanic	Nilo-Saharan	East Africa	Uganda	Medium
Makasae	Timor-Alor-Pantar	Trans-New Guinea	Australasia	East Timor	Medium
Malagasy	Barito	Austronesian	East Africa	Madagascar	Major
Mandarin	Sinitic	Sino-Tibetan	East Asia	China	Major
Manggarai	Sumba-Flores	Austronesian	Australasia	Indonesia	Medium
Mapudungun	Araucanian	Araucanian	South America	Chile	Medium
Maybrat	Bird’s Head	West Papuan	Australasia	Indonesia	Minor
Meithei	Manipuri	Sino-Tibetan	West Asia	India	Medium
Mien	Mienic	Hmong-Mien	East Asia	China	Medium
Miskito	Misumalpan	Misumalpan	North America	Nicaragua	Medium
Mixtec	Mixtecan	Oto-Manguean	North America	Mexico	Medium
Mon	Monic	Austro-Asiatic	East Asia	Thailand	Medium
Mongolian	Mongolic	Altaic	East Asia	Mongolia	Medium
Murrinhpatha	Murrinhpathan	Southern Daly	Australasia	Australia	Minor
Nahuatl	Aztecan	Uto-Aztecan	North America	Mexico	Medium
Navajo	South Athabaskan	Na-Dené	North America	United States	Medium
Nenets	Samoyedic	Uralic	West Asia	Russia	Minor
Newar	Himalayish	Sino-Tibetan	West Asia	Nepal	Medium
Nubian	Nubic	Nilo-Saharan	East Africa	Egypt	Medium
Nuosu	Loloish	Sino-Tibetan	East Asia	China	Medium
Otomi	Otomian	Oto-Manguean	North America	Mexico	Medium

Table A1. *Cont.*

Language	Genus	Family	Area	Country/Region	Size
Paez	Paezan	Paezan	South America	Colombia	Minor
Paiute	Numic	Uto-Aztecan	North America	United States	Minor
Paiwan	South Formosan	Austronesian	East Asia	Taiwan	Minor
Persian	Iranian	Indo-European	West Asia	Iran	Major
Pitjantjatjara	Wati	Pama-Nyungan	Australasia	Australia	Minor
Purepecha	Tarascan	Tarascan	North America	Mexico	Medium
Qaqet	Baining	East Papuan	Australasia	Papua New Guinea	Minor
Qiang	Qiangic	Sino-Tibetan	East Asia	China	Medium
Quechua	Quechuan	Andean	South America	Peru	Medium
Raramuri	Tarahumaran	Uto-Aztecan	North America	Mexico	Minor
Russian	Slavic	Indo-European	Europe	Russia	Major
Saami	Saamic	Uralic	Europe	Norway	Minor
Sandawe	Sandawan	Khoisan	East Africa	Tanzania	Minor
Sango	Ubangi	Niger-Congo	East Africa	Central Africa	Medium
Santali	Munda	Austro-Asiatic	West Asia	India	Medium
Savosavo	Central Solomon	East Papuan	Australasia	Solomon Islands	Minor
Shipibo	Panoan	Pano-Tacanan	South America	Peru	Minor
Sioux	Dakotan	Siouan	North America	United States	Minor
Slavey	North Athabaskan	Na-Dené	North America	Canada	Minor
Somali	East Cushitic	Afro-Asiatic	East Africa	Somalia	Major
Spanish	Romance	Indo-European	Europe	Spain	Major
Swahili	Bantu	Niger-Congo	East Africa	Tanzania	Major
Taa	Tuu	Khoisan	East Africa	Botswana	Minor
Tamil	South Dravidian	Dravidian	West Asia	India	Major
Telugu	Central Dravidian	Dravidian	West Asia	India	Major
Temne	Mel	Niger-Congo	West Africa	Sierra Leone	Medium
Ternate	Halmaheran	West Papuan	Australasia	Indonesia	Minor
Thai	Zhuang-Tai	Tai-Kadai	East Asia	Thailand	Major
Tibetan	Bodic	Sino-Tibetan	East Asia	China	Medium
Ticuna	Ticunan	Ticuna-Yuri	South America	Brazil	Minor
Tiwi	Tiwian	Tiwian	Australasia	Australia	Minor
Toba	Qom	Guaicuruan	South America	Argentina	Minor
Tok-Pisin	English-based	Creole	Australasia	Papua New Guinea	Medium
Totonac	Totonacan	Totonacan	North America	Mexico	Medium
Turkish	Oghuz Turkic	Altaic	West Asia	Turkey	Major
Udmurt	Permic	Uralic	West Asia	Russia	Medium
Uzbek	Karluk Turkic	Altaic	West Asia	Uzbekistan	Major
Vietnamese	Vietic	Austro-Asiatic	East Asia	Vietnam	Major
Wa	Palaungic	Austro-Asiatic	East Asia	Myanmar	Medium
Wandala	Central Chadic	Afro-Asiatic	West Africa	Cameroon	Medium
Warlpiri	Ngarrkic	Pama-Nyungan	Australasia	Australia	Minor
Wayuu	Goajiran	Arawakan	South America	Venezuela	Medium
Wichi	Wichi-Chorote	Matacoan	South America	Argentina	Minor
Wolaytta	Omotic	Afro-Asiatic	East Africa	Ethiopia	Medium
Xavante	Central Je	Macro-Je	South America	Brazil	Minor
Xun	Kxa	Khoisan	East Africa	Angola	Minor
Yakut	Siberian Turkic	Altaic	East Asia	Russia	Medium
Yanomami	Yanomaman	Yanomaman	South America	Brazil	Minor
Yoruba	Defoid	Niger-Congo	West Africa	Nigeria	Major
Yucatec	Yucatecan	Mayan	North America	Mexico	Medium
Zapotec	Zapotecan	Oto-Manguean	North America	Mexico	Medium
Zarma	Songhay	Nilo-Saharan	West Africa	Niger	Medium
Zoque	Zoquean	Mixe-Zoque	North America	Mexico	Medium

Appendix B

Table A2. Values of the phonological variables.

Language	Consonants	Vowels	VowQual	Tones	Stress	VowNasal	VowLength
Aguaruna	15	8	4	1	1	1	0
Akan	27	10	10	3	0	0	0
Albanian	29	7	7	1	0	0	0
Arabic	29	6	3	1	0	0	1
Armenian	30	6	6	1	0	0	0
Arrernte	27	4	4	1	1	0	0
Asheninka	23	8	4	1	0	0	1
Atayal	19	6	6	1	1	0	0
Aymara	26	6	3	1	0	0	1
Bambara	21	14	7	2	0	0	1
Basque	23	5	5	1	1	0	0
Batak	17	7	7	1	1	0	0
Beja	21	7	5	1	1	0	1
Berber	34	3	3	1	0	0	0
Blackfoot	12	6	3	2	0	0	1
Brahui	28	8	5	1	0	0	1
Bugis	19	6	6	1	0	0	0
Burmese	34	11	8	4	0	1	0
Burushaski	36	10	5	1	1	0	1
Cambodian	16	21	11	1	1	0	1
Chamorro	20	6	6	1	1	0	0
Chechen	38	10	5	1	0	0	1
Cherokee	23	11	6	6	0	0	1
Choctaw	16	9	3	1	1	1	1
Chukchi	14	3	3	1	0	0	0
Cree	10	7	4	1	0	0	1
Dani	13	14	7	1	1	0	1
Dholuo	26	9	9	3	0	0	0
Dogon	17	17	7	3	0	1	1
Embera	19	12	6	1	1	1	0
Enga	15	5	5	5	0	0	0
English	24	11	11	1	1	0	0
Evenki	18	13	7	1	0	0	1
Fijian	16	10	5	1	0	0	1
Filipino	16	5	5	1	1	0	0
Finnish	13	16	8	1	0	0	1
Fulfulde	27	7	7	1	0	0	0
Fur	17	8	8	2	0	0	0
Garifuna	17	6	6	1	1	0	0
Georgian	28	5	5	1	1	0	0
Greek	18	5	5	1	1	0	0
Guajajara	14	7	7	1	0	0	0
Guarani	18	12	6	1	1	1	0
Guaymi	25	16	8	1	1	1	0
Gumuz	36	10	5	2	0	0	1
Gunwinggu	22	5	5	1	0	0	0
Haitian	17	10	7	1	0	1	0
Hausa	28	10	5	2	0	0	1
Hindi	34	19	11	1	0	1	0

Table A2. *Cont.*

Language	Consonants	Vowels	VowQual	Tones	Stress	VowNasal	VowLength
Hlai	18	6	6	4	0	0	0
Hmong	58	8	8	7	0	0	0
Hungarian	25	14	7	1	0	0	1
Iatmul	21	12	7	1	0	0	1
Ibibio	13	12	7	2	0	0	1
Ijo	20	18	9	2	0	1	0
Indonesian	18	6	6	1	0	0	0
Inuit	14	6	3	1	0	0	1
Iraqw	29	10	5	2	0	0	1
Irish	35	11	11	1	0	0	0
Japanese	16	10	5	2	0	0	1
Jinghpo	31	10	5	4	0	0	1
Kabardian	53	3	2	1	1	0	1
Kabiye	21	9	9	2	0	0	0
Kam	27	6	6	10	0	0	0
Kamano	13	6	6	2	1	0	0
Kanuri	22	7	7	2	0	0	0
Karen	25	14	11	4	0	1	0
Kazakh	20	11	11	1	0	0	0
Khoekhoe	31	8	5	4	0	1	0
Kiche	22	10	5	1	0	0	1
Korean	19	18	9	1	0	0	1
Kuman	14	5	5	1	0	0	0
Kunama	22	10	5	3	0	0	1
Lezgian	54	6	6	1	1	0	0
Lithuanian	45	11	11	1	1	0	0
Macushi	10	12	6	1	1	0	1
Madi	45	9	9	3	0	0	0
Makasae	14	5	5	1	0	0	0
Malagasy	29	4	4	1	0	0	0
Mandarin	19	5	5	4	0	0	0
Manggarai	26	6	6	1	1	0	0
Mapudungun	22	6	6	1	0	0	0
Maybrat	11	5	5	1	1	0	0
Meithei	25	6	6	2	0	0	0
Mien	33	9	8	8	0	0	1
Miskito	14	6	3	2	0	0	1
Mixtec	16	10	6	3	0	1	0
Mon	27	10	10	2	0	0	0
Mongolian	26	14	7	1	0	0	1
Murrinhpatha	17	4	4	1	1	0	0
Nahuatl	15	8	4	1	1	0	1
Navajo	28	16	4	2	0	1	1
Nenets	27	9	6	1	0	0	1
Newar	26	20	6	1	0	1	1
Nubian	17	10	5	1	0	0	1
Nuosu	43	10	10	3	0	0	0
Otomi	23	12	8	3	0	1	0
Paez	35	16	4	1	1	1	1
Paiute	25	11	6	1	0	0	1
Paiwan	22	4	4	1	0	0	0
Persian	23	6	6	1	1	0	0

Table A2. *Cont.*

Language	Consonants	Vowels	VowQual	Tones	Stress	VowNasal	VowLength
Pitjantjatjara	17	6	3	1	0	0	1
Purepecha	25	6	6	1	1	0	0
Qaqet	16	7	4	1	0	0	1
Qiang	37	15	8	1	0	0	1
Quechua	25	3	3	1	0	0	0
Raramuri	19	5	5	3	1	0	0
Russian	36	6	6	1	1	0	0
Saami	35	10	5	1	0	0	1
Sandawe	44	15	5	2	0	1	1
Sango	26	12	7	3	0	1	0
Santali	21	14	8	1	0	1	0
Savosavo	17	5	5	1	1	0	0
Shipibo	15	8	4	1	1	1	0
Sioux	29	8	5	1	1	1	0
Slavey	36	15	5	2	0	1	1
Somali	22	10	10	3	0	0	0
Spanish	18	5	5	1	1	0	0
Swahili	32	5	5	1	0	0	0
Taa	87	28	5	2	0	1	1
Tamil	15	10	5	1	0	0	1
Telugu	35	12	6	1	0	0	1
Temne	19	9	9	2	0	0	0
Ternate	19	5	5	1	1	0	0
Thai	21	18	9	5	0	0	1
Tibetan	28	8	8	2	0	0	0
Ticuna	11	6	6	10	0	0	0
Tiwi	16	5	5	1	0	0	0
Toba	20	5	5	1	0	0	0
Tok-Pisin	17	5	5	1	1	0	0
Totonac	17	6	3	1	1	0	1
Turkish	22	8	8	1	0	0	0
Udmurt	26	7	7	1	1	0	0
Uzbek	26	6	6	1	0	0	0
Vietnamese	22	11	9	8	0	0	1
Wa	33	9	9	1	0	0	0
Wandala	41	3	3	2	0	0	0
Warlpiri	18	6	3	1	0	0	1
Wayuu	14	12	6	1	0	0	1
Wichi	34	5	5	1	1	0	0
Wolaytta	29	10	5	2	0	0	1
Xavante	13	13	9	1	0	1	0
Xun	94	21	5	4	0	1	1
Yakut	21	19	12	1	1	0	1
Yanomami	12	14	7	1	0	1	0
Yoruba	18	11	7	3	0	1	0
Yucatec	20	10	5	2	0	0	1
Zapotec	20	5	5	3	0	0	0
Zarma	20	16	5	4	0	1	1
Zoque	12	6	6	1	1	0	0

References

[1] Coloma, G., 2014. Towards a Synergetic Statistical Model of Language Phonology. Journal of Quantitative Linguistics. 21, 100–122. DOI: https://doi.org/10.1080/09296174.2014.882184

[2] Köhler, R., 2005. Synergetic Linguistics. In: Altmann, G., Köhler, R., Piotrowski, R. (eds.). Quantitative Linguistics: An International Handbook. De Gruyter: Berlin, Germany. pp. 760–774.

[3] Klymenko, O., Yenikeyeva, S., 2022. Synergetic Linguistics as a New Philosophy of Language Studies. Theory and Practice in Language Studies. 12, 417–423. DOI: https://doi.org/10.17507/tpls.1202.28

[4] Brentari, D., 2012. Phonology. In: Pfau, R., Steinbach, M., Woll, B. (eds.). Sign Language: An International Handbook. De Gruyter: Berlin, Germany. pp. 21–54.

[5] Hurford, J., 2014. The Origins of Language. Oxford University Press: Oxford, UK.

[6] Sundaram, R., 1996. A First Course in Optimization Theory. Cambridge University Press: Cambridge, UK.

[7] Dryer, M., Haspelmath, M., 2013. The World Atlas of Language Structures Online. Max Planck Institute for Evolutionary Anthropology: Leipzig, Germany.

[8] IPA, 1999. Handbook of the International Phonetic Association. Cambridge University Press: Cambridge, UK.

[9] Greene, W., 2020. Econometric Analysis, 8th ed. Pearson: Harlow, UK.

[10] Rasinger, S., 2013. Quantitative Research in Linguistics, 2nd ed. Bloomsbury: London, UK.

[11] Coloma, G., 2016. An Optimization Model of Global Language Complexity. Glottometrics. 35, 49–63.

[12] Coloma, G., 2017. Complexity Trade-Offs in the 100-Language WALS Sample. Language Sciences. 59, 148–158. DOI: https://doi.org/10.1016/j.langsci.2016.10.006

[13] Coloma, G., 2017. The Existence of Negative Correlation between Linguistic Measures across Languages. Corpus Linguistics and Linguistic Theory. 13, 1–26. DOI: https://doi.org/10.1515/cllt-2015-0020

[14] Coloma, G., 2022. Correlation between Linguistic Measures: An Extended Analysis. Studies in Linguistics and Literature. 6(4), 109–132. DOI: https://doi.org/10.22158/sll.v6n4p109