entropy

MDPI

Article

Syntactic Parameters and a Coding Theory Perspective on Entropy and Complexity of Language Families

Matilde Marcolli

Department of Mathematics, California Institute of Technology, Pasadena, CA 91125, USA; matilde@caltech.edu; Tel.: +1-626-395-4326

Academic Editors: Frédéric Barbaresco, Frank Nielsen and Kevin H. Knuth
Received: 14 January 2016; Accepted: 18 March 2016; Published: 7 April 2016

Abstract: We present a simple computational approach to assigning a measure of complexity and information/entropy to families of natural languages, based on syntactic parameters and the theory of error correcting codes. We associate to each language a binary string of syntactic parameters and to a language family a binary code, with code words the binary string associated to each language. We then evaluate the code parameters (rate and relative minimum distance) and the position of the parameters with respect to the asymptotic bound of error correcting codes and the Gilbert–Varshamov bound. These bounds are, respectively, related to the Kolmogorov complexity and the Shannon entropy of the code and this gives us a computationally simple way to obtain estimates on the complexity and information, not of individual languages but of language families. This notion of complexity is related, from the linguistic point of view to the degree of variability of syntactic parameter across languages belonging to the same (historical) family.

Keywords: syntax; principles and parameters; error-correcting codes; asymptotic bound; Kolmogorov complexity; Gilbert–Varshamov bound; Shannon entropy

1. Introduction

We propose an approach, based on Longobardi's parametric comparison method (PCM) and the theory of error-correcting codes, to a quantitative evaluation of the "complexity" of a language family. One associates to a collection of languages to be analyzed with the PCM a binary (or ternary) code with one code word for each language in the family and each word consisting of the binary values of the syntactic parameters of that language. The ternary case allows for an additional parameter state that takes into account certain phenomena of entailment of parameters. We then consider a different kind of parameters: the code parameters of the resulting code, which in coding theory account for the efficiency of the coding and decoding procedures. These can be compared with some classical bounds of coding theory: the asymptotic bound, the Gilbert–Varshamov (GV) bound, *etc*. The position of the code parameters with respect to some of these bounds provides quantitative information on the variability of syntactic parameters within and across historical-linguistic families. While computations carried out for languages belonging to the same historical family yield codes below the GV curve, comparisons across different historical families can give examples of isolated codes lying above the asymptotic bound.

1.1. Principles and Parameters

The generative approach to linguistics relies on the notion of a Universal Grammar (UG) and a related universal list of syntactic parameters. In the Principles and Parameters model, developed since [1], these are thought of as binary valued parameters or "switches" that set the grammatical structure

of a given language. Their universality makes it possible to obtain comparisons, at the syntactic level, between arbitrary pairs of natural languages.

A PCM was introduced in [2] as a quantitative method in historical linguistics, for comparison of languages within and across historical families at the syntactic instead of the lexical level. Evidence was given in [3,4] that the PCM gives reliable information on the phylogenetic tree of the family of Indo-European languages.

The PCM relies essentially on constructing a metric on a family of languages based on the relative Hamming distance between the sets of parameters as a measure of relatedness. The phylogenetic tree is then constructed on the basis of this datum of relative distances, see [3].

More work on syntactic phylogenetic reconstructions, involving a larger set of languages and parameters is ongoing, [5]. Syntactic parameters of world languages have also been used recently for investigations on the topology and geometry of syntactic structures and for statistical physics models of language evolution, [6–8].

Publicly available data of syntactic parameters of world languages can be obtained from databases such as Syntactic Structures of World Languages (SSWL) [9] or TerraLing [10] or World Atlas of Language Structures (WALS) [11]. The data of syntactic parameters used in the present paper are taken from Table A of [3].

1.2. Syntactic Parameters, Codes and Code Parameters

Our purpose in this paper is to connect the PCM approach to the mathematical theory of error-correcting codes. We associate a code to any group of languages one wishes to analyze via the PCM, which has one code word for each language. If one uses a number n of syntactic parameters, then the code C sits in the space $\mathbb{F}_2^n$, where the elements of $\mathbb{F}_2 = \{0, 1\}$ correspond to the two $\mp$ possible values of each parameter, and the code word of a language is the string of values of its n parameters. We also consider a version with codes on an alphabet $\mathbb{F}_3$ of three letters which allows for the possibility that some of the parameters may be made irrelevant by entailment from other parameters. In this case we use the letter $0 \in \mathbb{F}_3$ for the irrelevant parameters and the nonzero values ± 1 for the parameters that are set in the language.

In the theory of error-correcting codes, see [12], one assigns to a code $C \subset \mathbb{F}_q^n$ two code parameters: $R = \log_q(\#C)/n$, the transmission rate of the code, and $\delta = d/n$ the relative minimum distance of the code, where d is the miminum Hamming distance between pairs of distinct code words. It is well known in coding theory that "good codes" are those that maximize both parameters, compatibly with several constraints relating R and δ. Consider the function $f : \mathcal{C}_q \to [0, 1]^2$ from the space $\mathcal{C}_q$ of q-ary codes to the unit square, that assigns to a code C its code parameters, $f(C) = (\delta(C), R(C))$. A point (δ, R) in the range of f has finite (respectively, infinite) multiplicity if the preimage $f^{-1}(\delta, R)$ is a finite set (respectively, an infinite set). It was proved in [13] that there is a curve $R = \alpha_q(\delta)$ in the space of code parameters, the asymptotic bound, that separates code points that fill a dense region and that have infinite multiplicity from isolated code points that only have finite multiplicity. These better but more elusive codes are typically obtained through algebro-geometric constructions, see [13–15]. The asymptotic bound was related to Kolmogorov complexity in [16].

1.3. Position with Respect to the Asymptotic Bound

Given a collection of languages one wants to compare through their syntactic parameters, one can ask natural questions about the position of the resulting code in the space of code parameters and with respect to the asymptotic bound. The theory of error correcting codes tells us that codes above the asymptotic bound are very rare. Indeed, we considered various sets of languages, and for each choice of a set of languages we considered an associated code, with a code word for each language in the set, given by its list of syntactic parameters. When computing the code parameters of the resulting code, one finds that, in a range of cases we looked at, when the languages in the chosen set belong to the same historical-linguistic family the resulting code lies below the asymptotic bound (and in fact below

the Gilbert–Varshamov curve). This provides a precise quantitative bound to the possible spread of syntactic parameters compared to the size of the family, in terms of the number of different languages belonging to the same historico-linguistic group.

However, we also show that, if one considers sets of languages that do not belong to the same historical-linguistic family, then one can obtain codes that lie above the asymptotic bound, a fact that reflects, in code theoretic terms, the much greater variability of syntactic parameters. The result is in itself not surprising, but the point we wish to make is that the theory of error-correcting codes provides a natural setting where quantitative statements of this sort can be made using methods already developed for the different purposes of coding theory. We conclude by listing some new linguistic questions that arise by considering the parametric comparison method under this coding theory perspective.

1.4. Complexity of Languages and Language Families

The study of natural languages from the point of view of complexity theory has been of significant interest to linguists in recent years. The approaches typically followed focus on assigning a reasonable measure of complexity to individual languages and comparing complexities across different languages. For example, a notion of morphological complexity was studied in [17]. An approach to defining Kolmogorov complexity of languages on the basis of syntactic parameters was developed in [18]. A notion of language complexity based on the production rules of a generative grammar was considered in [19], in the setting of (finite) formal languages. For a more general computational perspective on the complexity of natural languages, see [20]. The idea of distinguishing languages by complexity is not without controversy in Linguistics. A very interesting general discussion of the problem and its evolution in the field can be found in [21].

In the present paper, we argue in favor of a somewhat different perspective, where we assign an estimate of complexity not to individual languages but to groups of languages, and in particular (historical) language families. Our version of complexity is measuring how "spread out" the syntactic parameters can be, across the languages that belong to the same family. As we outlined in the previous subsections, this is measured by assigning to the language family a code, whose code words record the syntactic parameters of the individual languages in the family, then computing its code parameters and evaluating the position of the resulting code points with respect to curves like the asymptotic bound or the Gilbert–Varshamov line. The reason why this position carries complexity information lies in the subtle relation between the asymptotic bound and Kolmogorov complexity, recently derived by Manin and the author in [16], which we will review briefly in this paper.

2. Language Families as Codes

The Principles and Parameters model of Linguistics assigns to every natural language L a set of binary values parameters that describe properties of the syntactic structure of the language.

Let F be a *language family*, by which we mean a finite collection $F = \{L_1, \dots, L_m\}$ of languages. This may coincide with a family in the historical sense, such as the Indo-European family, or a smaller subset of languages related by historic origin and development (e.g., the Indo-Iranian, or Balto–Svalic languages), or simply any collection of languages one is interested in comparing at the parametric level, even if they are spread across different historical families.

We denote by n be the number of parameters used in the parametric comparison method. We do not fix, a priori, a value for n, and we consider it a variable of the model. We will discuss below how one views, in our perspective, the issue of the independence of parameters.

After fixing an enumeration of the parameters, that is, a bijection between the set of parameters and the set $\{1, \dots, n\}$, we associate to a language family F a code $C = C(F)$ in $\mathbb{F}_2^n$, with one code word for each language $L \in F$, with the code word $w = w(L)$ given by the list of parameters $w = (x_1, \dots, x_n)$, $x_i \in \mathbb{F}_2$ of the language. For simplicity of notation, we just write L for the word $w(L)$ in the following.

In this model, we only consider binary parameters with values ± 1 (here identified with letters 0 or 1 in $\mathbb{F}_2$) and we ignore parameters in a neutralized state following implications across parameters, as in the datasets of [3,4]. The entailment of parameters, that is, the phenomenon by which a particular value of one parameter (but not the complementary value) renders another parameter irrelevant, was addressed in greater detail in [22]. We first discuss a version of our coding theory model that does not incorporate entailment. We will then comment in Section 2.7 below on how the model can be modified to incorporate this phenomenon.

The idea that natural languages can be described, at the level of their core grammatical structures, in terms of a string of binary characters (code words) was already used extensively in [23].

2.1. Code Parameters

In the theory of error-correcting codes, one assigns two main parameters to a code C, the *transmission rate* and the *relative minimum distance*. More precisely, a binary code $C \subset \mathbb{F}_2^n$ is an $[n, k, d]_2$-code if the number of code words is $\#C = 2^k$, that is,

$$k = \log_2 \#C, \tag{1}$$

where k need not be an integer, and the minimal Hamming distance between code words is

$$d = \min_{L_1 \neq L_2 \in C} d_H(L_1, L_2), \tag{2}$$

where the Hamming distance is given by

$$d_H(L_1, L_2) = \sum_{i=1}^{n} |x_i - y_i|,$$

for $L_1 = (x_i)_{i=1}^n$ and $L_2 = (y_i)_{i=1}^n$ in C. The transmission rate of the code C is given by

$$R = \frac{k}{n}. \tag{3}$$

One denotes by $\delta_H(L_1, L_2)$ the relative Hamming distance

$$\delta_H(L_1, L_2) = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|,$$

and one defines the relative minimum distance of the code C as

$$\delta = \frac{d}{n} = \min_{L_1 \neq L_2 \in C} \delta_H(L_1, L_2). \tag{4}$$

In coding theory, one would like to construct codes that simultaneously optimize both parameters (δ, R): a larger value of R represents a faster transmission rate (better encoding), and a larger value of δ represents the fact that code words are sufficiently sparse in the ambient space $\mathbb{F}_2^n$ (better decoding, with better error-correcting capability). Constraints on this optimization problem are expressed in the form of bounds in the space of (δ, R) parameters, see [12,13].

In our setting, the R parameter measures the ratio between the logarithmic size of the number of languages encompassing the given family and the total number of parameters, or equivalently how densely the given language family is in the ambient configuration space $\mathbb{F}_2^n$ of parameter possibilities. The parameter δ is the minimum, over all pairs of languages in the given family, of the relative Hamming distance used in the PCM method of [3,4].

2.2. Parameter Spoiling

In the theory of error-correcting codes, one considers *spoiling operations* on the code parameters. Applied to an $[n,k,d]_2$-code C, these produce, respectively, new codes with the following description (see Section 1.1.1 of [24]):

- A code $C_1 = C \star_i f$ in $\mathbb{F}_2^{n+1}$, for a map $f : C \to \mathbb{F}_2$, whose code words are of the form $(x_1,\dots,x_{i-1},f(x_1,\dots,x_n),x_i,\dots,x_n)$ for $w = (x_1,\dots,x_n) \in C$. If f is a constant function, C_1 is an $[n+1,k,d]_2$-code. If all pairs $w,w' \in C$ with $d_H(w,w') = d$ have $f(w) \neq f(w')$, then C_1 is an $[n+1,k,d+1]_2$-code.
- A code $C_2 = C\star_i$ in $\mathbb{F}_2^{n-1}$, whose code words are given by the projections

$$(x_1,\dots,x_{i-1},x_{i+1},\dots,x_n)$$

  of code words $(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n)$ in C. This is an $[n-1,k,d-1]_2$-code, except when all pairs $w,w' \in C$ with $d_H(w,w') = d$ have the same letter x_i, in which case it is an $[n-1,k,d]_2$-code.
- A code $C_3 = C(a,i) \subset C \subset \mathbb{F}_2^n$, given by the level set $C(a,i) = \{w = (x_k)_{k=1}^n \in C \mid x_i = a\}$. Taking $C(a,i)\star_i$ gives an $[n-1,k',d']_2$-code with $k-1 \leq k' < k$, and $d' \geq d$.

The same spoiling operations hold for q-ary codes $C \subset \mathbb{F}_q^n$, for any fixed q.

In our setting, where C is the code obtained from a family of languages, according to the procedure described above, the first spoiling operation can be seen as the effect of considering one more syntactic parameter, which is dependent on the other parameters, hence describing a function $F : \mathbb{F}_2^n \to \mathbb{F}_2$, whose restriction to C gives the function $f : C \to \mathbb{F}_2$. In particular, the case where f is constant on C represents the situation in which the new parameter adds no useful comparison information for the selected family of languages. The second spoiling operation consists in forgetting one of the parameters, and the third corresponds to forming subfamilies of the given family of languages, by grouping together those languages with a set value of one of the syntactic parameters. Thus, all these spoiling operations have a clear meaning from the point of view of the linguistic PCM.

2.3. Examples

We consider the same list of 63 parameters used in [3] (see Section 5.3.1 and Table A). This choice of parameters follows the *modularized global parameterization* method of [2], for the Determiner Phrase module. They encompass parameters dealing with person, number, and gender (1–6 on their list), parameters of definiteness (7–16 in their list), of countability (17–24), genitive structure (25–31), adjectival and relative modification (32–14), position and movement of the head noun (42–50), demonstratives and other determiners (51–50 and 60–63), possessive pronouns (56–59); see Section 5.3.1 and Section 5.3.2 of [3] for more details.

Our very simple examples here are just meant to clarify our notation: they consist of some collections of languages selected from the list of 28, mostly Indo-European, languages considered in [3]. In each group we consider we eliminate the parameters that are entailed from others, and we focus on a shorter list, among the remaining parameters, that will suffice to illustrate our viewpoint.

Example 1. Consider a code C formed out of the languages $\ell_1 =$ Italian, $\ell_2 =$ Spanish, and $\ell_3 =$ French, and let us consider only the first six syntactic parameters of Table A of [3], so that $C \subset \mathbb{F}_2^n$ with $n = 6$. The code words for the three languages are

ℓ_1	1	1	1	0	1	1
ℓ_2	1	1	1	1	1	1
ℓ_3	1	1	1	0	1	0

This has code parameters $(R = \log_2(3)/6 = 0.2642, \delta = 1/6)$, which satisfy $R < 1 - H_2(\delta)$, hence they lie below the GV curve (see Equation (8) below). We use this code to illustrate the three spoiling operations mentioned above.

- Throughout the entire set of 28 languages considered in [3], the first two parameters are set to the same value 1, hence for the purpose of comparative analysis within this family, we can regard a code like the above as a twice spoiled code $C = C' \star_1 f_1 = (C'' \star_2 f_2) \star_1 f_1$ where both f_1 and f_2 are constant equal to 1 and $C'' \subset \mathbb{F}_2^4$ is the code obtained from the above by canceling the first two letters in each code word.
- Conversely, we have $C'' = C' \star_2$ and $C' = C \star_1$, in terms of the second spoiling operation described above.
- To illustrate the third spoiling operation, one can see, for instance, that $C(0,4) = \{\ell_1, \ell_3\}$, while $C(1,6) = \{\ell_2, \ell_3\}$.

2.4. The Asymptotic Bound

The spoiling operations on codes were used in [13] to prove the existence of an *asymptotic bound* in the space of code parameters (δ, R), see also [16,24,25] for more detailed properties of the asymptotic bound.

Let $\mathcal{V}_q \subset [0,1]^2 \cap \mathbb{Q}^2$ denote the space of code parameters (δ, R) of codes $C \subset \mathbb{F}_q^n$ and let $\mathcal{U}_q$ be the set of all limit points of $\mathcal{V}_q$. The set $\mathcal{U}_q$ is characterized in [13] as

$$\mathcal{U}_q = \{(\delta, R) \in [0,1]^2 \mid R \leq \alpha_q(\delta)\}$$

for a continuous, monotonically decreasing function $\alpha_q(\delta)$ (the asymptotic bound). Moreover, code parameters lying in $\mathcal{U}_q$ are realized with infinite multiplicity, while code points in $\mathcal{V}_q \setminus (\mathcal{V}_q \cap \mathcal{U}_q)$ have finite multiplicity and correspond to the *isolated codes*, see [13,16].

Codes lying above the asymptotic bound are codes which have extremely good transmission rate and relative minimum distance, hence very desirable from the coding theory perspective. The fact that the corresponding code parameters are not limit points of other code parameters and only have finite multiplicity reflect the fact that such codes are very difficult to reach or approximate. Isolated codes are known to arise from algebro-geometric constructions, [14,15].

Relatively little is known about the asymptotic bound: the question of the computability of the function $\alpha_q(\delta)$ was recently addressed in [25] and the relation to Kolmogorov complexity was investigated in [16]. There are explicit upper and lower bounds for the function $\alpha_q(\delta)$, see [12], including the Plotkin bound

$$\alpha_q(\delta) = 0, \quad \text{for} \quad \delta \geq \frac{q-1}{q}; \tag{5}$$

the singleton bound, which implies that $R = \alpha_q(\delta)$ lies below the line $R + \delta = 1$; the Hamming bound

$$\alpha_q(\delta) \leq 1 - H_q(\frac{\delta}{2}), \tag{6}$$

where $H_q(x)$ is the q-ary Shannon entropy

$$x \log_q(q-1) - x \log_q(x) - (1-x) \log_q(1-x)$$

which is the usual Shannon entropy for $q = 2$,

$$H_2(x) = -x \log_2(x) - (1-x) \log_2(1-x). \tag{7}$$

One also has a lower bound given by the Gilbert–Varshamov bound

$$\alpha_q(\delta) \geq 1 - H_q(\delta) \tag{8}$$

The Gilbert–Varshamov curve can be characterized in terms of the behavior of sufficiently random codes, in the sense of the Shannon Random Code Ensemble, see [26,27], while the asymptotic bound can be characterized in terms of Kolmogorov complexity, see [16].

2.5. Code Parameters of Language Families

From the coding theory viewpoint, it is natural to ask whether there are codes C, formed out of a choice of a collection of natural languages and their syntactic parameters, whose code parameters lie above the asymptotic bound curve $R = \alpha_2(\delta)$.

For instance, a code C whose code parameters violate the Plotkin bound (5) must be an isolated code above the asymptotic bound. This means constructing a code C with $\delta \geq 1/2$, that is, such that any pair of code words $w \neq w' \in C$ differ by at least half of the parameters. A direct examination of the list of parameters in Table A of [3] and Figure 7 of [4] shows that it is very difficult to find, within the same historical linguistic family (e.g., the Indo-European family) pairs of languages L_1, L_2 with $\delta_H(L_1, L_2) \geq 1/2$. For example, among the syntactic relative distances listed in Figure 7 of [4] one finds only the pair (Farsi, Romanian) with a relative distance of 0.5. Other pairs come close to this value, for example Farsi and French have a relative distance of 0.483, but French and Romanian only differ by 0.162.

One has better chances of obtaining codes above the asymptotic bound if one compares languages that are not so closely related at the historical level.

Example 2. Consider the set $C = \{L_1, L_2, L_3\}$ with languages L_1 = Arabic, L_2 = Wolof, and L_3 = Basque. We exclude from the list of Table A of [3] all those parameters that are entailed and made irrelevant by some other parameter in at least one of these three chosen languages. This gives us a list of 25 remaining parameters, which are those numbered as 1–5, 7, 10, 20–21, 25, 27–29, 31–32, 34, 37, 42, 50–53, 55–57 in [3], and the following three code words:

L_1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	1	1	1	0	1	0	0	0	0
L_2	1	1	1	0	0	1	1	0	1	0	1	0	0	1	0	1	1	0	0	1	1	1	1	1	1
L_3	1	1	0	1	0	0	1	0	0	0	1	1	1	0	1	1	0	1	1	1	1	1	1	0	0

This example, although very simple and quite artificial in the choice of languages, already suffices to produce a code C that lies above the asymptotic bound. In fact, we have $d_H(L_1, L_2) = 16$, $d_H(L_2, L_3) = 13$ and $d_H(L_1, L_3) = 13$, so that $\delta = 0.52$. Since $R > 0$, the code point (δ, R) violates the Plotkin bound, hence it lies above the asymptotic bound.

It would be more interesting to find a code C consisting of languages belonging to the same historical-linguistic family (outside of the Indo-European group), that lies above the asymptotic bound. Such examples would correspond to linguistic families that exhibit a very strong variability of the syntactic parameters, in a way that is quantifiable through the properties of C as a code.

If one considers the 22 Indo-European languages in [3] with their parameters, one obtains a code C that is below the Gilbert–Varshamov line, hence below the asymptotic bound by Equation (8). A few other examples, taken from other non Indo-European historical-linguistic families, computed using those parameters reported in the SSWL database (for example the set of Malayo–Polynesian languages currently recorded in SSWL) also give codes whose code parameters lie below the Gilbert–Varshamov curve. One can conjecture that any code C constructed out of natural languages belonging to the same historical-linguistic family will be below the asymptotic bound (or perhaps below the GV bound), which would provide a quantitative bound on the possible spread of syntactic parameters within a historical family, given the size of the family. Examples like the simple one constructed above, using languages not belonging to the same historical family show that, to the contrary, across different historical families one encounters a greater variability of syntactic parameters. To our knowledge, no systematic study of parameter variability from this coding theory perspective has been implemented so far.

Ongoing work of the author is considering a systematic analysis of language families, based on the SSWL database of syntactic parameters, using this coding theory technique. This will include an analysis of how much the conclusions about the spreading of syntactic parameters across language families obtained with this technique depends on data pre-processing like the removal of spoiling features and what can be retained as an objective property of a set of languages. Moreover, a further purpose of this ongoing study is to combine the coding theory approach and the measures of complexity for groups of languages described in the present paper with the spin glass dynamical models of language change considered in [8], which was aimed at studying dynamically the spreading of syntactic parameters across groups of languages. The aim is to introduce complexity measures based on coding theory as part of the energy landscape of the spin glass model, following the suggestion of [28], on analogies between the roles of complexity in the theory of computation and energy in physical theories. These results, along with a more detailed analysis of the codes and code parameters of various language families, will appear in forthcoming work.

2.6. Comparison with Other Bounds

Another possible question one can consider in this setting is how the codes obtained from syntactic parameters of a given set of natural languages compare with other known families of error correcting codes and with other bounds in the space of code parameters.

For instance, it is known that an important improvement over the behavior of typical random codes can be obtained by considering codes determined by algebro-geometric curves defined over a finite field $\mathbb{F}_q$. Let $N_q(X) = \#X(\mathbb{F}_q)$ be the number of points over $\mathbb{F}_q$ of the curve X, and let $N_q(g) = \max N_q(X)$, with the maximum taken over all genus g curves X over $\mathbb{F}_q$. As shown in Theorem 2.3.22 of [12], asymptotically the $N_q(g)$ satisfy the Drinfeld–Vladut bound

$$A(q) := \limsup_{q\to\infty} \frac{N_q(g)}{g} \leq \sqrt{q} - 1,$$

and as shown in Section 3.4.1 of [12], this determines an algebro-geometric bound

$$\alpha_q(\delta) \geq R_{AG}(\delta) = 1 - \frac{1}{A(q)} - \delta$$

and the asymptotic Tsfasman–Vladut–Zink bound

$$\alpha_q(\delta) \geq R_{TVZ}(\delta) = 1 - (\sqrt{q} - 1)^{-1} - \delta.$$

The Tsfasman–Vladut–Zink line $R_{TVZ}(\delta) = 1 - (\sqrt{q} - 1)^{-1} - \delta$ lies entirely below the GV line for $q < 49$ (Theorem 3.4.4 of [12]).

A probabilistic argument given in Section 3.4.2 of [12] shows that highly non-random codes coming from algebraic curves can be asymptotically better than random codes (for sufficiently large q) as they cluster around the TVZ line. However, for $q = 2$ or $q = 3$, as in the case of codes from syntactic parameters of groups of languages that we consider here, the TVZ line lies below the GV line, hence any example that lies above the GV bound also behaves better than the the algebro-geometric bound. Such examples, like the one given above, for the three languages Arabic, Wolof, Basque, are very rare among codes obtained from syntactic parameters of languages, as they require the choice of a group of languages that are all very far from each other syntactically, with very large relative Hamming distances between syntactic parameters.

On the other hand, even for cases of groups of languages for which the resulting code parameters are below the GV line, it is still possible to get some additional information by comparing the position of the code parameters to other curves obtained from other bounds, such as the Blokh–Zyablow

bound or the Katsman–Tsfasman–Vladut bound, see Appendix A.2.1 of [12] for a summary of all these different bounds.

For example, the first example given above, with the three languages Italian, Spanish, French and a string of six syntactic parameters, gives a code with code parameters that are below the GV line, but above both the Blokh–Zyablow and the Katsman–Tsfasman–Vladut, according to the table of asymptotic bounds given in Appendix A.2.4 of [12].

2.7. Entailment and Dependency of Parameters

In the discussion above we did not incorporate in our model the fact that certain syntactic parameters can entail other parameters in such a way that one particular value of one of the parameters renders another parameter irrelevant or not defined, see the discussion in Section 5.3.2 of [3].

One possible way to alter the previous construction to account for these phenomena is to consider the codes C associated to families of languages as codes in $\mathbb{F}_3^n$, where n is the number of parameters, as before, and the set of values is now given by $\{-1, 0, +1\} = \mathbb{F}_3$, with ± 1 corresponding to the binary values of the parameters that are set for a given language and value 0 assigned to those parameters that are made irrelevant for the given language, by entailment from other parameters, or are not defined. This allows us to consider the full range of parameters used in [3,4]. We revisit Example 2 considered above.

Example 3. Let $C = \{L_1, L_2, L_3\}$ be the code obtained from the languages $L_1 =$ Arabic, $L_2 =$ Wolof, and $L_3 =$ Basque, as a code in $\mathbb{F}_3^n$ with $n = 63$, using the entire list of parameters in [3]. The code parameters $(R = 0.0252, \delta = 0.4643)$ of this code no longer violate the Plotkin bound. In fact, the parameters satisfy $R < 1 - H_3(\delta)$ so the code C now also lies below the GV bound.

Thus, the effect of including the entailed syntactic parameters in the comparison spoils the code parameters enough that they fall in the area below the GV bound.

Notice that what we propose here is different from the counting used in [3], where the relative distances $\delta_H(L_1, L_2)$ are normalized with respect to the number of non-zero parameters (which therefore varies with the choice of the pair (L_1, L_2)) rather than the total number n of parameters. While this has the desired effect of getting rid of insignificant parameters that spoil the code, it has the undesirable property of producing codes with code words of varying lengths, while counting only those parameters that have no zero-values over the entire family of languages, as in Example 2 avoids this problem. Adapting the coding theory results about the asymptotic bound to codes with words of variable length may be desirable for other reasons as well, but it will require an investigation beyond the scope of the present paper.

More generally, there are various kinds of dependencies among syntactic parameters. Some sets of hierarchical relations are discussed, for instance, in [29].

By the spoiling operations $C \star_i f$ of codes described above, we know that if some of the syntactic parameters considered are functions of other parameters, the resulting code parameters of $C \star_i f$ are worse than the parameters of the code C where only independent parameters were considered.

Part of the reason why code parameters of groups of languages in the family analyzed in [3] end up in the region below the asymptotic and the GV bound may be an artifact of the presence of dependences among the chosen 63 syntactic parameters. From the coding theory perspective, the parametric comparison method works best on a smaller set of independent parameters than on a larger set that includes several dependencies.

Entailment relations between syntactic parameters play an important role in the dynamical models of language evolutions constructed in [8], based on spin glass models in statistical physics.

Notice that the type of entailment relations we consider here are only of a rather special form, where a parameter is made undefined by effect of the value of another parameter (hence the use of the value 0 for the undetermined parameter). There are more general forms of entailment that we do not consider here, but which will be discussed in more detail in upcoming work. For example, one

can have a situation with two languages in which a parameter is entailed by the values of two other parameters, but entailed to two different values in the two languages. In this case, the proposal above need to be modified, because this entailed parameter should contribute to the Hamming distance between the two languages. In such a situation the entailed parameter should increase, rather than spoil, the efficiency of the code. Keeping entailed parameters can be used for error-correcting purposes, as contributing to error detection. The role of entailment of parameters was considered in [8], in the use of spin glass models for language change, where the entailment relations appear as couplings at the vertices (interaction terms) between different Ising/Potts models on the same underlying graph of language interactions. In upcoming work, now in preparation, we will discuss how treating different forms of entailment of parameters in the coding theory setting described here related to the treatment of entailment relations in the spin glass model of [8].

3. Entropy and Complexity for Language Families

3.1. Why the Asymptotic Bound?

In the examples discussed above we compared the position of the code point associated to a given set of languages to certain curves in the space of code parameters. In particular, we focused on the asymptotic bound curve and the Gilbert–Varshamov curve. It should be pointed out that these two curves have a very different nature.

The asymptotic bound is the only curve that separates regions in the space of parameters that correspond to code points with entirely different behavior. As shown in [13,24], code points in the area below the asymptotic bound are realized with infinite multiplicity and fill densely the region, while code points that lie above the asymptotic bound are isolated and realized with finite multiplicity.

The Gilbert–Varshamov curve, by contrast, is related to the statistical behavior of sufficiently random codes (as we recall in Section 3.2 below), but does not separate two regions with significantly different behavior in the space of code points. Thus, in this respect, the asymptotic bound is a more natural curve to consider than the Gilbert–Varshamov curve.

Thus, a heuristic interpretation of the position of codes obtained from groups of languages, with respect to the asymptotic bound can be understood as follows. The position of a code point above or below the asymptotic bound reflects a very different behavior of the corresponding code with respect to how easily "deformable" it is. The sporadic codes that lie above the asymptotic bound are rigid objects, in contrast to the deformable objects below the asymptotic bound. In terms of properties of the distribution of syntactic parameters within a set of languages, this different nature of the associated code can be seen as a measure of the degree of "deformability" of the parameter distribution: in languages that belong to the same historical linguistic families, the parameter distribution has evolved historically along with the development of the family's phylogenetic tree, and one expects that correspondingly the code parameters will indicate a higher degree of "deformability" of the corresponding code. If a group of languages is chosen that belong to very different historical families, on the contrary, one expects that the distribution of syntactic parameters will not necessarily lead any longer to a code that has the same kind of deformability property: code points above the asymptotic bound may be realizable by this type of language groups.

There is no similar interpretation for the position of the code point with respect to the Gilbert–Varshamov line. An interpretation of that position can be sought in terms of Shannon entropy, as we discuss below. Summarizing: the main conceptual distinction between the Gilbert–Varshamov line and the asymptotic bound is that the GV line represents only a statistical phenomenon, as we review below, while the asymptotic bound represents a true separation between two classes of structurally different codes, in the sense explained above.

3.2. Entropy and Statistics of the Gilbert–Varshamov Line

The Gilbert–Varshamov line $R = 1 - H_q(\delta)$ can be characterized statisticallly. Such a statistical description can be obtained by considering the Shannon Random Code Ensemble (SRCE). These are random codes obtained by choosing code words as independent random variables with respect to a uniform Bernoulli measure, so that a code is described by a randomly chosen set of different words of length n occurring with probability q^{-n}, see [26,27]. There is no a priori reason why the type of codes we consider here, with code words formed using the syntactic parameters of natural languages, would be linear. Thus, we consider the general setting of unstructured codes, as in Section V of [27].

The Hamming volume $Vol_q(n, d = n\delta)$, that is, the number of words of length n at Hamming distance at most d from a given one, can be estimated in terms of the q-ary Shannon entropy

$$H_q(\delta) = \delta \log_q(q-1) - \delta \log_q \delta - (1-\delta) \log_q(1-\delta)$$

in the form

$$q^{(H_q(\delta)-o(1))n} \leq Vol_q(n, d = n\delta) = \sum_{j=0}^{d} \binom{n}{j} (q-1)^j \leq q^{H_q(\delta)n}.$$

The expectation value for the random variable counting the number of unordered pairs of distinct code words with Hamming distance at most d is then estimated as

$$\mathbb{E} \sim \binom{q^k}{2} Vol_q(n, d) q^{-n} \sim q^{n(H_q(\delta)-1+2R)+o(n)}.$$

This estimate is then used (see [26,27]) to show that the probability to have codes in the SRCE with $H_q(\delta) \geq \max\{1-2R, 0\} + \epsilon$ is bounded by $q^{-\epsilon n(1+o(1))}$. By a similar argument (see Section V of [27] and Proposition 2.2 of [16]) it is shown that the probability that $H_q(\delta) \geq 1 - R + \epsilon$ is bounded by $q^{-n\epsilon(1+o(1))}$.

While, by this type of argument, one can see the Gilbert–Varshamov line as representing the typical behavior of sufficiently random codes, the asymptotic bound does not have a similar statistical interpretation. It does have, however, a relation to Kolmogorov complexity, which is relevant to the point of view discussed in the present paper. The relation between asymptotic bound of error correcting codes and Kolmogorov complexity was described in [16]. We recall it in the rest of this section, along with its implications for the linguistic applications we are considering.

3.3. Kolmogorov Complexity

We refer the reader to [30] for an extensive treatment of Kolmogorov complexity and its properties. We recall here some basic facts we need in the following.

Let $T_{\mathcal{U}}$ be a universal Turing machine, that is, a Turing machine that can simulate any other arbitrary Turing machine, by reading on tape both the input and the description of the Turing machine it should simulate. A prefix Turing machine is a Turing machine with unidirectional input and output tapes and bidirectional work tapes. The set of programs P on which a prefix Turing machine halts forms a prefix code.

Given a string w in an alphabet $\mathfrak{A}$, the prefix Kolmogorov complexity is given by minimal length of a program for which the universal prefix Turing machine $T_{\mathcal{U}}$ outputs w,

$$\mathcal{K}_{T_{\mathcal{U}}}(w) = \min_{P : T_{\mathcal{U}}(P) = w} \ell(P).$$

There is a universality property. Namely, given any other prefix Turing machine T, one has

$$\mathcal{K}_T(w) \leq \mathcal{K}_{T_{\mathcal{U}}}(w) + c_T,$$

where the shift is by a bounded constant, independent of w. The constant c_T is the Kolmogorov complexity of the program needed to describe T so that $T_{\mathcal{U}}$ can simulate it.

A variant of the notion of Kolmogorov complexity described above is given by conditional Kolmogorov complexity,

$$\mathcal{K}_{T_{\mathcal{U}}}(w \,|\, \ell(w)) = \min_{P:T_{\mathcal{U}}(P,\ell(w))=w} \ell(P),$$

where the length $\ell(w)$ is given, and made available to the machine $T_{\mathcal{U}}$. One then has

$$\mathcal{K}(w \,|\, \ell(w)) \leq \ell(w) + c,$$

because if $\ell(w)$ is known, then a possible program is just to write out w. This means that then $\ell(w) + c$ is just number of bits needed for the transmission of w plus the print instructions.

An upper bound is given by

$$\mathcal{K}_{T_{\mathcal{U}}}(w) \leq \mathcal{K}_{T_{\mathcal{U}}}(w \,|\, \ell(w)) + 2\log \ell(w) + c.$$

If one does not know a priori $\ell(w)$, one needs to signal the end of the description of w. For this it suffices to have a "punctuation method", and one can see that this has the effect of adds the term $2\log \ell(w)$ in the above estimate. In particular, any program that produces a description of w is an upper bound on Kolmogorov complexity $\mathcal{K}_{T_{\mathcal{U}}}(w)$.

One can think of Kolmogorov complexity in terms of data compression: the shortest description of w is also its most compressed form. Upper bounds for Kolmogorov complexity are therefore provided easily by data compression algorithms. However, while providing upper bounds for complexity is straightforward, the situation with lower bounds is entirely different: constructing a lower bound runs into a fundamental obstacle caused by the fact that the halting problem is unsolvable. As a consequence, Kolmogorov complexity is not a computable function. Indeed, suppose one would list programs P_k (with increasing lengths) and run them through the machine $T_{\mathcal{U}}$. If the machine halts on P_k with output w, then $\ell(P_k)$ is an approximation to $\mathcal{K}_{T_{\mathcal{U}}}(w)$. However, there may be an earlier P_j in the list such that $T_{\mathcal{U}}$ has not yet halted on P_j. If $T_{\mathcal{U}}$ eventually halts also on P_j and outputs w, then $\ell(P_j)$ will be a better approximation to $\mathcal{K}_{T_{\mathcal{U}}}(w)$. So one would be able to compute $\mathcal{K}_{T_{\mathcal{U}}}(w)$ if one could tell exactly on which programs P_k the machine $T_{\mathcal{U}}$ halts, but that is indeed the unsolvable halting problem.

Kolmogorov complexity and Shannon entropy are related: one can view Shannon entropy as an averaged version of Kolmogorov complexity in the following sense (see Section 2.3 of [31]). Suppose given independent random variables X_k, distributed according to Bernoulli measure $\mathbb{P} = \{p_a\}_{a\in\mathfrak{A}}$ with $p_a = \mathbb{P}(X = a)$. The Shannon entropy is given by

$$S(X) = -\sum_{a\in\mathfrak{A}} \mathbb{P}(X = a)\log \mathbb{P}(X = a).$$

There exists a $c > 0$, such that, for all $n \in \mathbb{N}$,

$$S(X) \leq \frac{1}{n}\sum_{w\in\mathcal{W}^n} \mathbb{P}(w)\,\mathcal{K}(w \,|\, \ell(w)) \leq S(X) + \frac{\#\mathfrak{A}\,\log n}{n} + \frac{c}{n}.$$

The expectation value

$$\lim_{n\to\infty} \mathbb{E}(\frac{1}{n}\mathcal{K}(X_1 \cdots X_n \,|\, n)) = S(X)$$

shows that the average expected Kolmogorov complexity for length n descriptions approaches the Shannon entropy in the limit when $n \to \infty$.

3.4. Kolmogorov Complexity and the Asymptotic Bound

We recall here briefly the result of [16] linking the asymptotic bound of error correcting codes to Kolmogorov complexity.

As we discussed above, only the asymptotic bound marks a significant change of behavior of codes across the curve (isolated code points with finite multiplicity versus accumulation points with infinite multiplicity). In this sense this curve is very different from all the other bounds in the space of code parameters. However, there is no explicit expression for the curve $R = \alpha_q(\delta)$ that gives the asymptotic bound. Indeed, even the question of the computability of the function $R = \alpha_q(\delta)$ is a priori unclear. This question was formulated explicitly in [25].

It is proved in [16] that the asymptotic bound $R = \alpha_q(\delta)$ becomes computable given an oracle that can list codes by increasing Kolmogorov complexity. Given such an oracle, one can provide an explicit iterative (algorithmic) procedure for constructing the asymptotic bound. This implies that the asymptotic bound is "at worst as non-computable as Kolmogorov complexity".

Consider the set $X = \mathcal{C}_q$ of (unstructured) q-ary codes and the set $Y \subset [0,1]^2$ of code points and the computable function $f : X \to Y$ that assigns to a code $C \in X$ its code parameters $(R(C), \delta(C)) \in Y$. Let Y_{fin} and Y_∞ be, respectively, the subsets of the space of code points that correspond to code points realized with finite and with infinite multiplicity. The algorithm iteratively produces two sets A_m and B_m that approximate, respectively, Y_∞ and Y_{fin} by $Y_{fin} = \cup_{m\geq 1} B_m$ and $Y_\infty = \cup_{m\geq 1}(\cap_{n\geq 0} A_{m+n})$. The inductive construction starts by choosing an increasing sequence of positive integers N_m and setting $B_1 = \emptyset$ and taking A_1 to be the set of code points y with $\nu_Y^{-1}(y) \leq N_1$, where $\nu_Y : \mathbb{N} \to Y$ is a fixed enumeration of the set of rational points $[0,1]^2 \cap \mathbb{Q}^2$ where code points belong.

General estimates on the behavior of (exponential) Kolmogorov complexity under composition of total recursive functions (see [30], Section VI.9 of [32]) show that, for a composition $F = f_0(f_1(t_1, \dots, t_m), \cdots, f_n(t_1, \dots, t_m), t_{m+1}, \dots, t_\ell)$ of recursive functions the Kolmogorov complexity satisfies

$$\mathcal{K}(F) \leq c \cdot \prod_{i=1}^{n} \mathcal{K}(f_i) \cdot \left(\log \prod_{i=1}^{n} \mathcal{K}(f_i) \right)^{n-1},$$

for a fixed f_0 and varying f_i, $i \geq 1$.

Consider the total recursive function $F(x) = (f(x), n(x))$ with

$$n(x) = \#\{x' \,|\, \nu_X^{-1}(x') \leq \nu_X^{-1}(x),\, f(x') = f(x)\}$$

where $\nu_X : \mathbb{N} \to X$ is an enumeration of the space of codes. Consider the enumerable sets $X_m := \{x \in X \,|\, n(x) = m\}$ and $Y_m := f(X_m) \subset Y$, with $Y_\infty = \cap_m f(X_m)$ and $Y_{fin} = f(X) \smallsetminus Y_\infty$. For $\varphi : f(X) \to X_1$, defined as f^{-1} on $f(X_1) = f(X)$, applying the composition rule for exponential Kolmogorov complexity, it is shown in Proposition 3.1 of [16] that, for $x \in X_1$ and $y = f(x)$, one has $\mathcal{K}(x) = \mathcal{K}(\varphi(y)) \leq c_\varphi \cdot \mathcal{K}(y) \leq c\nu_Y^{-1}(y)$, hence

$$\mathcal{K}_{T_\mathcal{U}}(x) \leq c \cdot \nu_Y^{-1}(y).$$

Using the same type of estimate of Kolmogorov complexity for composition of recursive functions, it is then shown in Proposition 3.2 [16] that, for $y \in Y_\infty$ and $m \geq 1$, and for a unique $x_m \in X$, with $y = f(x_m)$, $n(x_m) = m$ and $c = c(f, u, v, \nu_X, \nu_Y) > 0$, one finds

$$\mathcal{K}_{T_\mathcal{U}}(x_m) \leq c \cdot \nu_Y^{-1}(y)\, m\, \log(\nu_Y^{-1}(y) m).$$

To construct inductively A_{m+1} and B_{m+1}, given A_m and B_m, one takes A_{m+1} to consist of the elements in the list

$$\mathcal{L}_{m+1} = \{y \in f(X) \,:\, \nu_Y^{-1}(y) \leq N_{m+1},\ \exists x \in X, \text{ with } y = f(x) \text{ and } n(x) = m+1\}.$$

Here one invokes the oracle, which ensures that, if such x exists, then it must be contained in a finite list of points $x \in X$ with bounded complexity

$$\mathcal{K}_{T_{\mathcal{U}}}(x_m) \leq c \cdot \nu_Y^{-1}(y)\, m\, \log(\nu_Y^{-1}(y) m).$$

One then takes B_{m+1} to consist of the remaining elements in the list $\mathcal{L}_{m+1}$. We refer the reader to [16] for a more detailed formulation.

More generally, the argument of [16] recalled above shows that, for a recursive function $f : \mathbb{Z}_+ \to \mathbb{Q}$, determining which values have infinite multiplicities is computable given an oracle that enumerate integers in order of Kolmogorov complexity.

As discussed in [16,24], the asymptotic bound can also be seen as the phase transition curve for a quantum statistical mechanical system constructed out of the space of codes, where the partition function of the system weights codes according to their Kolmogorov complexity. This is as close to a "statistical description" of the asymptotic bound that one can achieve.

In comparison with the behavior of random codes (codes whose complexity is comparable to their size), which concentrate in the region bounded by the Gilbert–Varshamov line, when ordering codes by complexity, non-random codes of lower complexity populate the region above, with code points accumulating in the intermediate region bounded by the asymptotic bound. That intermediate region thus, in a sense, reflects the difference between Shannon entropy and complexity.

3.5. Entropy and Complexity Estimates for Language Families

On the basis of the considerations of the previous sections and of the results of [16,24] recalled above, we propose a way to assign a quantitative estimate of entropy and complexity to a given set of natural languages.

As before let $C = \{L_1, \dots, L_k\}$ be a binary (or ternary) code where the code words L_i are the binary (ternary) strings of syntactic parameters of a set of languages L_i. We define the *entropy* of the language family $\{L_1, \dots, L_k\}$ as the q-ary Shannon entropy $H_q(\delta(C))$, where q is either 2 or 3 for binary or ternary codes, and $\delta(C)$ is the relative minimum distance parameter of the code C. We also define the *entropy gap* of the language family $\{L_1, \dots, L_k\}$ as the value of $H_q(\delta(C)) - 1 + R(C)$, which measures the distance of the code point $(R(C), \delta(C))$ from the Gilbert–Varshamov line, that is, from the behavior of a typical random code.

As a source of estimates of complexity of a language family $\{L_1, \dots, L_k\}$ one can consider any upper bound on Kolmogorov complexity of the code C. A possible approach, which contains more linguistic input, would be to provide estimates of complexity for each individual language in the family and then compare these. Estimates of complexity for individual languages have been considered in the literature, some of them based on the description of languages in terms of their syntactic parameters. For instance, following [18], for a syntactic parameter Π with possible values $v \in \{\pm 1\}$, the Kolmogorov complexity of Π set to value v is given by

$$\mathcal{K}(\Pi = v) = \min_{\tau \text{ expressing } \Pi} \mathcal{K}_{T_{\mathcal{U}}}(\tau),$$

with the minimum taken over the complexities of all the parse trees that express the syntactic parameter Π and require $\Pi = v$ to be grammatical in the language. Notice that, in this approach, the syntactic parameters are not just regarded as binary or ternary values, but one needs to consider actual parse trees of sentences in the language that express the parameter. Thus, such an approach to complexity

has the advantage that it is very rich in linguistic information. However, it is at the same time computationally very difficult to realize.

What we are proposing here is a much simpler way to obtain an estimate of complexity for a language family $\{L_1, \ldots, L_k\}$, which is not based on estimating complexity of the individual languages in the family, but which is aimed at detecting how spread out and diversified the syntactic parameters are across the family, by estimating the position of the code point $(R(C), \delta(C))$ of the associated code C with respect to the asymptotic bound $R = \alpha_q(\delta)$. This can be estimated in terms of the distance to other curves in the space of code parameters (R, δ) that constrain the asymptotic bound from above and below, such as the Plotkin bound, Hamming bound, and Gilbert–Varshamov bound, as in the examples discussed in the previous sections.

4. Conclusions

We proposed an approach to estimating entropy and complexity of groups of natural languages (language families), based on the linguistic parametric comparison method (PCM) of [2,22] via the mathematical theory of error-correcting codes, by assigning a code to a family of languages to be analyzed with the PCM, and investigating its position in the space of code parameters, with respect to the asymptotic bound and the GV bound. We have shown that there are examples of languages not belonging to the same historical-linguistic family that yield isolated codes above the asymptotic bound, while languages belonging to the same historical-linguistic family appear to give rise to codes below the bound, though a more systematic analysis would be needed to map code parameters of different language groups. We have also shown that, from these coding theory perspective, it is preferable to exclude from the PCM all those parameters that are entailed and made irrelevant by other parameters, as those spoil the properties of the resulting code and produce code parameters that are artificially low with respect to the asymptotic bound and the GV bound.

The approach proposed here, based on the PCM and the theory of error-correcting codes, suggests a few new linguistic questions that may be suitable for treatment with coding theory methods:

1. Do languages belonging to the same historical-linguistic family always yield codes below the asymptotic bound or the GV bound? How often does the same happen across different linguistic families? How much can code parameters be improved by eliminating spoiling effects caused by dependencies and entailment of syntactic parameters?
2. Codes near the GV curve are typically coming from the Shannon Random Code Ensemble, where code words and letters of code words behave like independent random variables, see [26,27]. Are there families of languages whose associated codes are located near the GV bound? Do their syntactic parameters mimic the uniform Poisson distribution of random codes?
3. The asymptotic bound for error-correcting codes was related in [16] to Kolmogorov complexity, and the measure of complexity for language families that we proposed here is estimated in terms of the position of the code point with respect to the asymptotic bound. There are other notions of complexity, most notably the type of organized complexities discussed in [33–35]. Can these be related to loci in the space of code parameters? What do these represent when applied to codes obtained from syntactic parameters of a set of natural languages?
4. Is there a more direct linguistic complexity measure associated to a family of natural languages that would relate to the position of the resulting code above or below the asymptotic bound?
5. Codes and the asymptotic bound in the space of code parameters were recently studied using methods from quantum statistical mechanics, operator algebra and fractal geometry, [24,36]. Can some of these mathematical methods be employed in the linguistic parametric comparison method?

The observational results reported here are still preliminary. The following topics should be consolidated:

- How much the conclusions obtained for a given family of languages will depend on data pre-processing (removal of "spoiling" features, *etc*.)
- To what extent the proposed criterion (above or below the asymptotic bound) is really an objective property of a set of languages.

This will be addressed more thoroughly in future work. The concern about the effect of data pre-processing in paticular requires more analysis, that will be developed in further ongoing work, as outlined at the end of Section 2.5.

Acknowledgments: The author's research is supported by NSF grants DMS-1201512 and PHY-1205440, and by the Perimeter Institute for Theoretical Physics. The author thanks the referees for their useful comments.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Chomsky, N. *Lectures on Government and Binding*; Foris: Dordrecht, The Netherlands, 1981.
2. Longobardi, G. Methods in parametric linguistics and cognitive history. *Linguist. Var. Yearb.* **2003**, *3*, 101–138.
3. Longobardi, G.; Guardiano, C. Evidence for syntax as a signal of historical relatedness. *Lingua* **2009**, *119*, 1679–1706.
4. Longobardi, G.; Guardiano, C.; Silvestri, G.; Boattini, A.; Ceolin, A. Toward a syntactic phylogeny of modern Indo-European languages. *J. Hist. Linguist.* **2013**, *3*, 122–152.
5. Aziz, S.; Huynh, V.L.; Warrick, D.; Marcolli, M. Syntactic Phylogenetic Trees. 2016, In Preparation.
6. Park, J.J.; Boettcher, R.; Zhao, A.; Mun, A.; Yuh, K.; Kumar, V.; Marcolli, M. Prevalence and recoverability of syntactic parameters in sparse distributed memories. 2015, arXiv:1510.06342.
7. Port, A.; Gheorghita, I.; Guth, D.; Clark, J.M.; Liang, C.; Dasu, S.; Marcolli, M. Persistent Topology of Syntax. 2015, arXiv:1507.05134.
8. Siva, K.; Tao, J.; Marcolli, M. Spin Glass Models of Syntax and Language Evolution. 2015, arXiv:1508.00504.
9. Syntactic Structures of the World's Languages (SSWL) Database of Syntactic Parameters. Available online: http://sswl.railsplayground.net (accessed on 18 March 2016).
10. TerraLing. Available online: http://www.terraling.com (accessed on 18 March 2016).
11. Haspelmath, M.; Dryer, M.S.; Gil, D.; Comrie, B. *The World Atlas of Language Structures*; Oxford University Press: Oxford, UK, 2005.
12. Tsfasman, M.A.; Vladut, S.G. Algebraic-Geometric Codes. In *Mathematics and Its Applications (Soviet Series)*; Springer: Amsterdam, the Netherlands, 1991; Volume 58.
13. Manin, Y.I. What is the maximum number of points on a curve over $\mathbb{F}_2$? *J. Fac. Sci. Univ. Tokyo Sect. 1A Math.* **1982**, *28*, 715–720.
14. Tsfasman, M.A.; Vladut, S.G.; Zink, T. Modular curves, Shimura curves, and Goppa codes, better than Varshamov–Gilbert bound. *Math. Nachr.* **1982**, *109*, 21–28.
15. Vladut, S.G.; Drinfel'd, V.G. Number of points of an algebraic curve. *Funct. Anal. Appl.* **1983**, *17*, 68–69.
16. Manin, Y.I.; Marcolli, M. Kolmogorov complexity and the asymptotic bound for error-correcting codes. *J. Differ. Geom.* **2014**, *97*, 91–108.
17. Bane, M. Quantifying and measuring morphological complexity. In Proceedings of the 26th West Coast Conference on Formal Linguistics, Berkeley, CA, USA, 27–29 April 2007.
18. Clark, R. *Kolmogorov Complexity and the Information Content of Parameters*; Institute for Research in Cognitive Science: Philadelphia, PA, USA, 1994.
19. Tuza, Z. On the context-free production complexity of finite languages. *Discret. Appl. Math.* **1987**, *18*, 293–304.
20. Barton, G.E.; Berwick, R.C.; Ristad, E.S. *Computational Complexity and Natural Language*; MIT Press: Cambrige, MA, USA, 1987.
21. Sampson, G.; Gil, D.; Trudgill, P. (Eds.) *Language Complexity as an Evolving Variable*; Oxford University Press: Oxford, UK, 2009.
22. Longobardi, G. A minimalist program for parametric linguistics? In *Organizing Grammar: Linguistic Studies in Honor of Henk van Riemsdijk*; Broekhuis, H.; Corver, N.; Huybregts, M.; Kleinhenz, U.; Koster, J., Eds.; Mouton de Gruyter: Berlin, Germany, 2005; pp. 407–414.

23. Clark, R.; Roberts, I. A computational model of language learnability and language change. *Linguist. Inq.* **1993**, *24*, 299–345.
24. Manin, Y.I.; Marcolli, M. Error-correcting codes and phase transitions. *Math. Comput. Sci.* **2001**, *5*, 133–170.
25. Manin, Y.I. A computability challenge: Asymptotic bounds and isolated error-correcting codes. 2011, arXiv:1107.4246.
26. Barg, A.; Forney, G.D. Random codes: minimum distances and error exponents. *IEEE Trans. Inf. Theory* **2002**, *48*, 2568–2573.
27. Coffey, J.T.; Goodman, R.M. Any code of which we cannot think is good. *IEEE Trans. Inf. Theory* **1990**, *36*, 1453–1461.
28. Manin, Y.I. Complexity vs Energy: Theory of Computation and Theoretical Physics. 2014, arXiv:1302.6695.
29. Baker, M.C. *The Atoms of Language: The Mind's Hidden Rules of Grammar*; Basic Books: New York, NY, USA, 2001.
30. Li, M.; Vitányi, P. *An Introduction to Kolmogorov Complexity and Its Applications*; Springer: New York, NY, USA, 2008.
31. Grünwald, P.; Vitányi, P. Shannon Information and Kolmogorov Complexity. 2004, arXiv:cs/0410002.
32. Manin, Y.I. *A Course in Mathematical Logic for Mathematicians*, 2nd ed; Springer: New York, NY, USA, 2010.
33. Bennett, C.; Gacs, P.; Li, M.; Vitanyi, P.; Zurek, W. Information distance. *IEEE Trans. Inf. Theory* **1998**, *44*, 1407–1423.
34. Delahaye, J.P. *Complexité Aléatoire et Complexité Organisée*; Éditions Quæ: Paris, France, 2009. (In French)
35. Gell-Mann, M.; Lloyd, S. Information measures, effective complexity, and total information. *Complexity* **1996**, *2*, 44–52.
36. Marcolli, M.; Perez, C. Codes as fractals and noncommutative spaces. *Math. Comput. Sci.* **2012**, *6*, 199–215.