The Mirage of Morphological Complexity

Fermín Moscoso del Prado Martín (fermosc@gmail.com)
Laboratoire Dynamique du Langage, CNRS & Université de Lyon, Lyon, France &
Institut Rhône-Alpin des Systèmes Complexes (IXXI), Lyon, France

Abstract

I present a study on the morphological complexity of six European languages. A theory-free measure of the complexity of a language's inflectional morphology, is derived from Gell-Mann's concept of Effective Complexity. Using a parallel corpus, I show that disconsidering word order information results in the classical gradation of inflectional complexity: Languages in which words have more inflected variants seem to be more complex than languages with fewer variants. It also appears that the presence of the inflectional system increases the complexity of languages. However, when word order information is explicitly considered, the apparent gradation in complexity across languages vanishes. Furthermore, it becomes clear that the presence of inflections reduces the overall complexity of languages. In sum, inflection is used whenever its presence simplifies a language's description. Inflectional morphology is not a capricious feature, as some argue, but rather an effective tool for complexity reduction.

Keywords: Effective Complexity, Entropy, Morphology

Introduction

Words are usually accepted as being the fundamental units of syntax. However, in many (or most) languages, words can be related to each other by (quasi-)regular relations between their orthographical forms, their grammatical functions, and their meanings. These intra-lexical regularities are referred to as morphological structure, and constitute one of the main areas in the study of human languages, both from a linguistic and a psychological perspective.

Morphology: 'By itself'?

The classical structuralist position on inflectional systems is that they are largely random classifications that need not serve any useful function within the overall linguistic system. This is still a common view both among linguists (e.g., Aronoff, 1994; Bloomfield, 1933) and psychologists (e.g., Maratsos, 1979). For instance, Aronoff (1994) argues that morphology must be considered 'by itself', not merely from the perspective of its interface with syntax, phonology, or semantics, and that linguistic theory should therefore consider a separate and autonomous morphological component. In contrast, supporters of 'functionalist' theories of grammar would argue that the presence of inflectional systems must be useful from a communicative perspective. Otherwise, in line with Zipf (1949)'s 'Principle of Least Effort', the cognitive system would have discarded them for requiring unnecessary efforts. From this perspective, the presence of inflectional morphology is only justified if one or both of the following conditions apply:

1. Inflectional morphology expresses semantic content that is required for successful communication.
2. The presence of inflectional categories and markers reduces the overall complexity of a language.

What is morphological complexity?

Notice that the discussion above assumes that some objective quantification of the complexity of a morphological system is available. However, no definition or quantification of the complexity of a linguistic system is widely accepted. The concept of complexity has central importance for theories of language structure and processing. For instance, McWhorter (2001) raised an interesting controversy over his claim that pidgin and creole languages are intrinsically less complex than other languages, reflecting the very early stage in their development. However –despite its frequent invocation– linguistic complexity remains a rather ill-defined concept. Indeed, much previous research has been aimed at providing a concise and workable definition of linguistic –and morphological– complexity (e.g., Goldsmith, 2001; Greenberg, 1954; Juola, 1998, 2007; Malvern, Richards, Chipere, & Durán, 2004; McWhorter, 2001; Nichols, 2010; Shosted, 2006; Siegel, 2004; Vulanović, 2007; Xanthos & Gillis, 2010). Many of these studies define complexity in terms of either counting forms, or detecting the presence or absence of certain morphological phenomena or structures that are chosen as representative of complexity based on some theoretical premises, an approach which was also used by Lupyan and Dale (2010). However, such theory-bound measures can result in considerable vagueness and subjectivity with regard to what is more or less complex (see, e.g., McWhorter, 2001; Nichols, 2010; Shosted, 2006, for debate on this issue). To obtain more objective estimates, some have measured the complexity of inflectional paradigms as the number of inflected variants that can be formed for a single word (e.g. Malvern et al., 2004; Xanthos & Gillis, 2010). This simple approach presents the problem that it ignores that there are sometimes very regular relations between different inflected forms (e.g., the plural of an English word is in most cases the same as the singular plus an 's'), and not much information is required to encode this (what some term the Paradigm Cell Filling Problem; cf., Ackerman, Blevins, & Malouf, 2009). To avoid this, some have taken a theory-free approach investigating to what extent linguistic utterances or structures can be compressed (Goldsmith, 2001; Juola, 1998, 2007). The irreducible information that is left after compression provides a useful index of the informational content of the signal; its Algorithmic Information Content (AIC; Chaitin, 1987; Kolmogorov, 1965), and can thus be taken as a theory-free measure of complexity.[1] Notice that, by themselves, AIC approaches based on the compressibility of a corpus (such as Juola, 1998, 2007) do not in fact constitute measures of complexity; it is easy to see that AIC would rate an incompressible totally random corpus as being more complex than the complete works of Shakespeare, as the latter can be compressed to some degree. In contrast, the approach of Goldsmith (2001) overcomes this limitation. It does not measure the compressibility of the corpus, but rather, the compressibility of its formal regularities (a random text would have no regularities to compress). However, by summarizing paradigms as sets of forms and affixes, Goldsmith (2001)'s approach overlooks the crucial role that can be played by the actual functions served by each inflected variant, as these relate to the complexity of the system, and have been shown to be of crucial importance for the cognitive system (Kostić, Marković, & Baucal, 2003).

Outline

In what follows, I start by using the definition of the Effective Complexity of language (Moscoso del Prado Martín, submitted) to derive a measure of inflectional complexity. This is followed by a corpus-based analysis of the inflectional complexity of six European languages, investigating the variation in morphological complexity according to different measures. I conclude by discussing the theoretical implications of the results.

Formulation & Computational Methods

Effective Complexity of Language

As in Moscoso del Prado Martín (submitted), the definition of linguistic complexity is derived from Gell-Mann (1995)'s general definition of *Effective Complexity*:

> A measure that corresponds [...] to [...] complexity [...] refers not to the length of the most concise description of an entity (which is roughly what AIC is), but to the length of a concise description of a set of the entity's regularities. Thus something almost entirely random, with practically no regularities, would have effective complexity near zero. So would something completely regular, such as a bit string consisting entirely of zeroes.

For human languages, such descriptions of the system's regularities are grammars. Indeed, many have advocated that the best measure of a language's complexity would be the size of the shortest grammar that could fully describe it (Goldsmith, 2001; McWhorter, 2001). As a starting point, let us assume that the regularities that need to be accounted for correspond to all the sentences that appear in an arbitrarily large reference corpus of a language. This is to say, the definition of grammatical complexity rests on a reference corpus containing N characters (or phonemes, *etc.*), and corresponds to the length of the shortest possible grammar that can generate all the sentences in that corpus (*i.e.*, it is complete) and only those (*i.e.*, it does not over-generate). Notice this definition leaves open the grammatical theory or formalism in which the grammar is expressed: Any formal mechanism that is able to generate the sentences in a language is valid candidate. Let us denote the length of that optimal grammar as $G(N)$. In parallel to the definition above, one can also consider the AIC of the reference corpus, that is, the length of the shortest possible algorithmic description enabling its full reconstruction and denote the length of that optimal compression by $H(N)$.

On the one hand, the grammar that determines $G(N)$ needs to be able to generate all sentences in the corpus. On the other hand, the compressed version of the corpus also needs to generate all those very sentences, with only the additional burden of having to reconstruct their actual ordering and frequencies of occurrence. As both $H(N)$ and $G(N)$ are defined in terms of ideal 'optimal' methods that do not waste any space, one can decompose

$$H(N) = G(N) + H_s(N), \tag{1}$$

where $H_s(N) \geq 0$ denotes the additional information that needs to be coded in the AIC. It is useful to think of $G(N)$ and $H(N)$ in terms of per-character rates; their relation to the size of the reference corpus:

$$g(N) = \frac{1}{N} G(N) = \\ = \frac{1}{N}\left[H(N) - H_s(N)\right] = h(N) - h_s(N). \tag{2}$$

If a finite grammar for a language does exist, then, for increasingly large corpora, the grammar should come closer to being 'complete'. That is to say, from some large N onwards, $G(N)$ should grow much more slowly than N itself. One can now take the idealization a step further, and require that the ideal grammar be able to generate all sentences (and only those) that could *eventually* happen in the language (i.e., they have non-zero probabilities of occurrence). This is equivalent to taking the limit of an infinite corpus size. In this way, one defines the *grammatical complexity of the language* as

$$G = \lim_{N \to \infty} G(N). \tag{3}$$

At this extreme, the finite grammar should be complete, hence its size would become negligible compared to that

[1]The approach advocated by Goldsmith (2001) is not completely theory-free, as it requires the parsing of words into stems and affixes, which is in itself a strong theoretical commitment.

of the corpus. Defining the *grammatical density of the language* as the infinite corpus size limit of $g(N)$,

$$g = \lim_{N\to\infty} g(N) = \lim_{N\to\infty} \frac{G(N)}{N}. \quad (4)$$

If G is finite, one should find that

$$g = \lim_{N\to\infty} \frac{G(N)}{N} = 0. \quad (5)$$

That is to say, for a language to have a finite grammar, its grammatical density should be zero so that the limits in Eqs. 3–5 converge. One can also extend this limiting condition to the compressibility measures, and write

$$g = \lim_{N\to\infty} g(N) = \lim_{N\to\infty} [h(N) - h_s(N)] = h - h_s = 0. \quad (6)$$

On the one hand, h reflects the rate of compressibility of the original corpus, taken to the limit of infinite corpus size. On the other hand, h_s is the rate of compressibility of a manipulated version of the corpus, where the sentence identities, frequencies, and orderings are maintained, but their actual internal structure is lost. For stationary[2] ergodic sources, it is guaranteed that h and h_s are actually the source entropies (Shannon, 1948) of the original and modified versions of the corpus (Chaitin, 1987; Kolmogorov, 1965).

Moscoso del Prado Martín (submitted) found that, for human languages $g > 0$, and therefore no finite grammar can ever be found that fully accounts for all sentences in the language without generating impossible sentences. In other words, the grammatical complexity G of languages is not finite, but rather keeps growing for a growing reference corpus. In contrast, the grammatical density measure g provides a finite value that can be compared across languages, corresponding to the average information that is provided by each new character or phoneme in the large corpus size limit. This is to say, even if the actual grammatical complexity is not finite, one can compare the speed at which the complexity increases with increasing corpus size. Different languages can be encoded with more or less transparent orthographies, or may make use of longer, but very predictable expressions. Therefore, it is important to consider the grammatical density, not in terms of characters or phonemes, but using instead the basic unit that is being considered for the grammar, that is, the sentence. We can therefore define the per-sentence grammatical density of the language as:

$$g_s = L_s \cdot g, \quad (7)$$

[2]Stationarity is not a property of linguistic corpora, but it can be enforced by simply randomizing the order of the sentences in the corpus (Moscoso del Prado Martín, submitted). This discards all supra-sentential information, but we are assuming that the grammar must produce all well-formed sentences, and information beyond the sentence is thus not relevant here.

where L_s denotes the mean length of a sentence in whichever units (characters, phonemes, . . .) g was computed. This new measure provides a finite index of the complexity of a language, precisely measuring how much grammatical knowledge is provided by each new observed sentence.

Inflectional complexity

Assuming that one can somehow separate the different contributions to the per-sentence grammatical density that are provided by the different components of the grammar (below the sentence level) one could decompose g_s into something like

$$g_s = g_s^{\text{lexicon}} + g^{\text{inflection}} + g_s^{\text{derivation}} + g_s^{\text{syntax}} + \ldots . \quad (8)$$

Consider now that one could erase from the corpus all inflectional information without disturbing any of the other levels. In this case, one would obtain a new version of the grammatical densities that would discount all inflectional information,

$$g'_s = g_s'^{\text{lexicon}} + g_s'^{\text{derivation}} + g_s'^{\text{syntax}} + \ldots, \quad (9)$$

such that one can write,

$$g_s^{\text{inflection}} = g_s - g'_s. \quad (10)$$

I refer to $g_s^{\text{inflection}}$ as the *inflectional complexity* of a language, and it measures the average amount of information required to describe the new inflectional structures contained by a newly observed sentence. This measure is readily computable, and it enables direct comparison of the inflectional systems of different languages. Notice also that its actual values are themselves meaningful. A positive $g_s^{\text{inflection}}$ indicates that the presence of the morphological system increases the size of the grammar that is required to describe the language. On the other hand, were $g_s^{\text{inflection}}$ to be negative, it would indicate that the presence of the inflectional structures in fact *simplifies* the grammar of the language; the grammar would be more complex if the inflectional structures were absent.

Computations

The discussion above assumes that the entropy measures h and h_S can be accurately estimated from corpora. Here, I provide a summary of the methods used for estimating these magnitudes, see Moscoso del Prado Martín (submitted) for a more detailed discussion of these methods and their accuracy.

Estimation of h: Asymptotic Lempel-Ziv complexity The source entropy h indexes the compressibility of the corpus. Although knowing what is the shortest possible (the most compressed) version of a sequence is impossible unless one explicitly knows the dynamics of the process that generated it, some specific lossless compression algorithms are guaranteed to converge to this maximum in the long length limit. Of these, a particularly simple choice is to use the Lempel-Ziv algorithm (LZ76; Lempel & Ziv, 1976). For very long sequences, the Lempel-Ziv complexity almost surely converges to the entropy of the source (Ziv & Lempel, 1978). Hence, the LZ76 algorithm can be used for estimating the source entropies h. However, one should take into account that, although the LZ complexity asymptotically approaches h, estimates of h based directly on some large corpus are still subject to finite sample effects. To a large degree however, the finite sample error in estimation can be corrected for by modelling the convergence of the Lempel-Ziv measure. In this respect, Schürmann and Grassberger (1996) found that the convergence of the Lempel-Ziv measure ($L[n]$) as a function of the corpus size n can be very accurately modelled by a three parameter power-law expression, and one of these parameters corresponds to the asymptotic value of the entropy h. These three parameters can be fitted from data. Given a finite size corpus, one can compute the Lempel-Ziv complexity using different subsets of the corpus of increasing length, and fit a simple non-linear regression provides the estimate of h. This approximation has been found to perform extremely well (Moscoso del Prado Martín, submitted; Schürmann & Grassberger, 1996).

Estimation of h_s: Chao-Shen estimator The other entropy rate factor, h_S corresponds to the per-character entropy rate of a sequence which preserves the identity and ordering of the sentences, but lacks their internal structure. A simple way to estimate this is to consider the corpus as a sequence of N_s symbols $S_1\,S_2\ldots S_{N_s}$, each of which corresponds to a full sentence. Note here that the alphabet on which this sequence is defined is unknown and possibly not even finite. Having such a large, perhaps infinite alphabet, makes the use of the Lempel-Ziv method inappropriate. For the LZ76 algorithm to begin to converge towards the source entropy, one needs to have observed at least a great proportion of the symbols in the original alphabet. However, even in extremely large corpora, a great majority of the sentences will occur exactly once, and a great many more possible sentences will not be in the corpus at all. On the other hand, as the sequence $S_1\,S_2\ldots S_{N_S}$ originates from a corpus in which the order of the sentences has been randomized, it is clear that the sequence is an ensemble of N_S independent and identically distributed random variables. To solve such type of problems, Chao and Shen (2003) developed an estimator that takes into account the contribution of unseen items to the overall entropy. The Chao-Shen estimator performs remarkably well in sequences originating from infinite alphabets (cf., Moscoso del Prado Martín, submitted). The estimator relies on a Good-Turing (Good, 1953) adjustment of the frequencies of occurrence of the sentences. The entropy is then computed on the adjusted values of the sentence frequencies, and the obtained estimator is further adjusted to account for the yet unseen sentences. Although the Chao-Shen estimate converges relatively rapidly to the true entropy value, in this case it still tends to underestimate the true value of the entropy. However, Moscoso del Prado Martín (submitted) found that an asymptotic correction of the Schürmann and Grassberger (1996) type is very accurate in this case as well.

Materials and Methods

I used the Dutch, English, French, German, Italian, and Spanish sections of the *Europarl Corpus* (Koehn, 2005), version 5. This corpus contains transcripts of sessions of the European Parliament between 1998 and 2009, transcribed (some partly) in eleven of the official languages of the E.U. For each language, the sessions of the Parliament were collapsed into eleven files, each corresponding to one individual year, and all computations were repeated for each year, as to provide multiple estimates for each language, all equivalent across the six languages. Within each of the eleven files in each language, the order of the sentences was randomized to ensure stationarity.

In order to provide versions of the corpora that preserved all linguistic information except for that corresponding to inflectional structures, I also created new versions of each corpora that had been tokenized, that is, all inflected forms were replaced by their base forms (i.e., “cars” was replaced by “car”, “ate” was replaced by “eat”, etc.). Notice that this manipulation preserves all orthographic, lexical, syntactic, and semantic structure, but lacks only the inflectional details. The tokenizing was performed using the *TreeTagger* software (Schmid, 1994) with the available scripts for each of the languages.[3]

Finally, to investigate the effect that functional information about syntax and semantics has on inflectional complexity, additional versions of all corpora (original and tokenized) were created by randomizing the order of the words (including sentence breaks) in each corpus. This produces versions of the corpora which retain most lexical and ortho/phonological information as well as the same diversity and probability distribution of inflectional forms that was present in the originals. The measures of inflectional complexity obtained from these randomized corpora therefore consider the information about diversity of inflectional forms, the regularity of their formation, and the homogeneity of their probability distributions. They do not consider any information regarding the functions that each inflectional form serves in the text, or the particular cell in the inflectional paradigms they occupy.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Results and Discussion

Figure 1 summarizes the resulting inflectional complexity values that were obtained for the different corpora. The upper panel plots the distribution of complexity values for the corpora in which the word order was left intact. Notice that, in this case, there was no significant heterogeneity in the distributions of complexity across the six languages compared. All complexity values seem to arise from a single distribution. More importantly, the inflectional complexity values tended to be *negative* for all languages but Dutch, for which it was not significantly different than zero. In other words, considering word order information, the inflectional morphology systems in fact simplify the grammar of the languages.

In contrast, the lower panel plots the estimated distributions of inflectional complexity when the word order information has been disregarded. In this case, there is indeed significant heterogeneity in the complexity of the different languages, and this variation fits well with the intuitions that one would obtain from form-counting measures of inflectional complexity: English $\leq$ Dutch $<$ German $\leq$ Romance Languages. Whereas English and Dutch have very limited inflectional systems, German makes use of a relatively complex nominal declension system, and Romance languages have rather complex verbal conjugation mechanisms including tens of forms. Furthermore, if word order is disconsidered, the complexity values are in all cases significantly positive, this is to say, the inclusion of the inflectional system makes the description of the languages more complex. Notice, however, that this differences in morphological complexity are but a mirage produced by not considering the actual functions served by each form.

I have introduced an objective and theoretically motivated measure of inflectional complexity. This new measure overcomes all the problems that were detailed in the introduction: it considers the diversity of forms, the regularity of their relations, the position that each form occupies in the paradigm, the frequency distribution of forms, and the function that each forms serves. Furthermore, the measure relies on an adequate concept of complexity, in contrast with the plain randomness that is reflected by raw measures of corpus compressibility. Finally, by extrapolating to infinite corpus size, one attenuates the innacuracies that are introduced by making inferences on finite-sized corpora.

These results highlight the strong degree of mutual dependence between morphological and syntactic information. In line with the behavioral results of Kostić et al. (2003), the structure of the morphological system cannot be considered independently of the actual functions that the different forms serve. In fact, considering the presence of these functions, inflectional morphology serves a role in reduction of uncertainty, simplifying the description of the whole grammar. Therefore, the presence of inflectional morphology does not seem to be in any way capricious, but instead it reflects a delicately optimized system.

Acknowledgments

I am indebted to A. Boussidan for proof-reading this manuscript.

References

Ackerman, F., Blevins, J. P., & Malouf, R. (2009). Parts and wholes: Implicative patterns in inflectional paradigms. In J. P. Blevins & J. Blevins (Eds.), *Analogy in grammar: Form and acquisition* (pp. 54–81). Oxford, UK: Oxford University Press.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Bloomfield, L. (1933). *Language*. New York, NY: Holt.

Chaitin, J. G. (1987). *Algorithmic information theory*. Cambridge, England: Cambridge University Press.

Chao, A., & Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species. *Environmental and Ecological Statistics*, *10*, 429–443.

Gell-Mann, M. (1995). What is complexity? *Complexity*, *1*, 16-19.

Goldsmith, J. A. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*, 153–198.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Greenberg, J. (1954). A quantitative approach to morphological typology of language. In R. Spencer (Ed.), *Method and perspective in anthropology* (pp. 192–195). Minneapolis, MN: University of Minnesota Press.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitive Linguistics*, *5*, 206–213.

Juola, P. (2007). Assessing linguistic complexity. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 89–108). Amsterdam, The Netherlands: John Benjamins.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT summit X* (pp. 79–86). Phuket, Thailand: Asia-Pacific Association for Machine Translation. Available from `http://www.statmt.org/europarl/`

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, *1*, 4–7.

Kostić, A., Marković, T., & Baucal, A. (2003). Inflectional morphology and word meaning: orthogonal or co-implicative domains? In R. H. Baayen & R. Schreuder (Eds.), *Morphological structure in lan-*

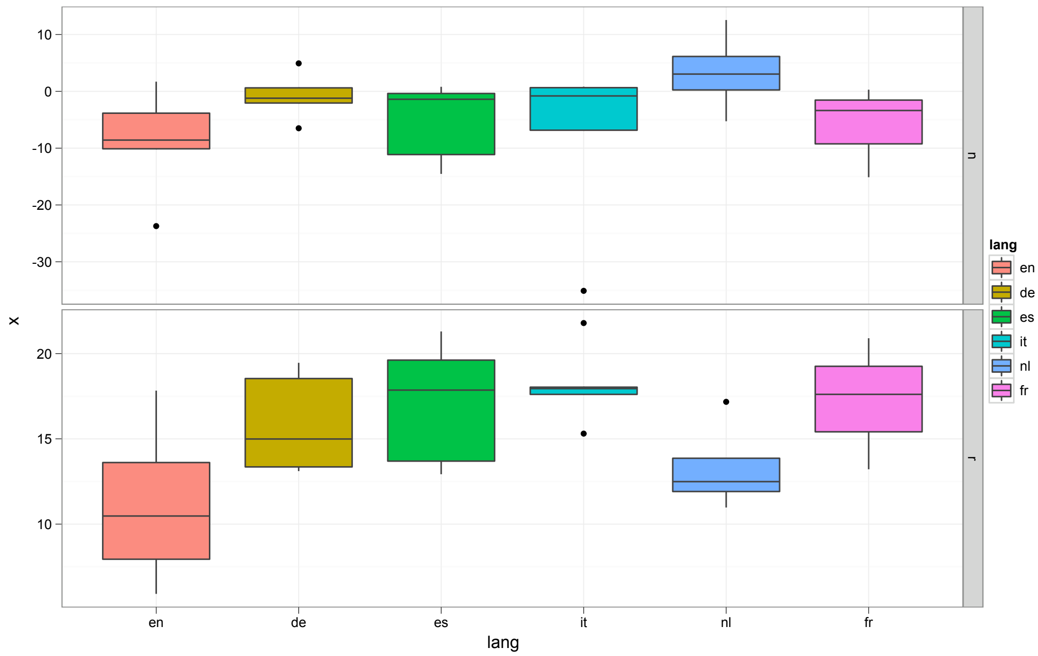

Figure 1: Summary of results. The upper panel plots the distribution of inflectional complexity (in nats/sentence) values obtained for each language in the original word order corpora. The lower panel plots the same results for the corpora in which the word order was randomized.

guage processing (pp. 1–44). Berlin, Germany: Mouton de Gruyter.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions in Information Theory*, *22*, 75–81.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS ONE*, *5*, e8559.

Malvern, D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment.* Palgrave Macmillan.

Maratsos, M. (1979). Learning how and when to use pronouns and determiners. In P. Fletcher & M. Garman (Eds.), *Language acquisition* (pp. 225–240). Cambridge, UK: Cambridge University Press.

McWhorter, J. H. (2001). The world's simplest grammars are creole grammars. *Linguistic Typology*, *5*, 125–166. (With replies and rejoinder)

Moscoso del Prado Martín, F. (submitted). *The Effective Complexity of language: English requires at least an infinite grammar.* Available from `http://moscosodelprado.net/docs/complexity-cognt.pdf`

Nichols, J. (2010). Linguistic complexity: a comprehensive definition and survey. In G. B. Sampson, D. Gil, & P. Trudgill (Eds.), *Language complexity as an evolving variable* (pp. 109–124). Oxford, England: Oxford University Press.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing.* Manchester, UK.

Schürmann, T., & Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos*, *6*, 414–427.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379-423, 623-656.

Shosted, R. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, *10*, 1–40.

Siegel, J. (2004). Morphological simplicity in pidgins and creoles. *Journal of Pidgin and Creole Languages*, *19*, 139–162.

Vulanović, R. (2007). On measuring language complexity as relative to the conveyed linguistic information. *SKY Journal of Linguistics*, *20*, 399–427.

Xanthos, A., & Gillis, S. (2010). Quantifying the development of inflectional diversity. *First Language*, *30*, 175 –198.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Cambridge, MA: Addison Wesley.

Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable length coding. *IEEE Transactions in Information Theory*, *24*, 530–536.