Manuscript submitted to *Cognitive Science* (please do not cite yet)

Cognitive Implications of Multifractal Structure in Language: A theoretical study and an empirical investigation

Fermín MOSCOSO DEL PRADO MARTÍN
Laboratoire Dynamique du Language (UMR–5596)
CNRS & Université de Lyon II, Lyon, France &
Institut Rhône-Alpin de Systèmes Complexes, France

Abstract

This study presents a detailed information theoretical analysis of human language. In particular, I investigate the implications that long-range correlations and multifractal scaling found in language have for issues as the rate at which information is produced or processed in language, and in relation to the growingly popular hypothesis stating that this information remains on average constant throughout discourse. I analyze several corpora across languages and genres. The results indicate that language exhibits multifractal structure, with heterogeneous fractal properties across different time scales. Importantly, I find that there is a direct correspondence between the scaling regimes and traditional linguistic levels of study. The findings support the adjustments of the information processing rate to optimize communication, rather than just signal transmission. This implies that the information processing rate does not remain constant, but most likely decreases along time. I show that the online adjustments lead to an intrinsic non-stationarity of linguistic sequences, which has important cognitive implications. Finally, I conclude by detailing how the results observed fit into the known facts from the psycholinguistic literature, and how they provide new insights into human language processing.

Keywords: Constant Information Flow; Entropy; Language Complexity; Long-Range Correlations; Multifractal Scaling; Uniform Information Density; Symbolic Dynamics

The amount of information that is conveyed by linguistic utterances is an old question, dating back to the groundbreaking work of Shannon (1951). This question has important implications in several fields, including telecommunications, the study of the structure of language itself (linguistics), and the study of how the human mind processes language (psycholinguistics). In recent years, it has been hypothesized that the rate at which linguistic information gets processed during

Correspondence can be addressed to:
fermosc@gmail.com

language production or language comprehension remains roughly constant. Therefore, as language is but a by-product of the cognitive system, the information (in the sense of Shannon, 1948) that is supplied by actual linguistic streams must as well be about constant in time. This has been referred to in various domains as the Constant Entropy Rate hypothesis (CER; Genzel & Charniak, 2002, 2003; Keller, 2004; Manin, 2006; Qian & Jaeger, 2009; Vega & Ward, 2009), the Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004), the Constant Information Flow (Fenk-Oczlon & Fenk, 1999, 2002, 2007), or the Uniform Information Density (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007; Piantadosi, Tily, & Gibson, 2009) hypothesis.

In general, it is found that elements in a linguistic signal tend to provide information about future elements to appear, and this predictivity seems to extend over relatively long lags (Cover & King, 1978; Shannon, 1951). More strongly, it appears that these correlations may extend long enough to consider the possibility of infinite correlation lengths such as those found in stochastic fractals, in the sense of Mandelbrot and Van Ness (1968). This possibility has attracted attention from physical scientists, and a considerable number of studies have been dedicated to exploring the fractal properties of linguistic signals and sequences using techniques from statistical physics (Amit, Shmerler, Eisenberg, Abraham, & Shnerb, 1994; Dębowski, 2006; Ebeling, 1997; Ebeling & Frömmel, 1998; Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Gillet & Ausloos, 2008; Hilberg, 1990; Kello, Anderson, Holden, & Van Orden, 2008; Kokol & Podgorelec, 2000; Lardner et al., 1992; Li, 1989; Maragos & Potamianos, 1999; Sabanal & Nakagawa, 1996; Schenkel, Zhang, & Zhang, 1993; Pavlov, Ebeling, Molgedey, Ziganshin, & Anishchenko, 2001; Pickover & Khorasani, 1986; Voss & Clarke, 1975). It is important to realize that, were this fractal structure present, it would have implications on the plausibility of the CER hypothesis mentioned above.

One of the rare points on which there is an almost unanimous agreement across the Linguistics community is that linguistic utterances are the result of multiple levels of structure at different scales (phonological, lexical, syntactic, *etc.*), and the regularities across these scales need not share the same properties. This kind of 'cascaded' multilevel structure is also found in many natural structures and, when combined with heterogeneous – but still fractal – scaling properties at each of the levels, it gives rise to the so-called multifractal systems (Ivanov et al., 1999; Stanley & Meakin, 1988). Indeed, the degree of heterogeinity that is found in texts is suggestive of such a multifractal structure (Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Gillet & Ausloos, 2008; Lardner et al., 1992; Pavlov et al., 2001).

In this study, I combine techniques from information theory and statistical physics to investigate the properties of human language from an information processing perspective. I pay special attention to the importance of long-range correlations and multifractal structure, and their direct relation with levels of linguistic description, and with psychological theories on how humans process language. In what follows, I introduce in more detail the CER hypothesis, together with the linguistic and psycholinguistic evidence that has been put forward in its support, and I will discuss some important problems with the theory. This is followed by a theoretical section in which I introduce several concepts from information theory and fractal scaling, as well as some theoretical derivations on the relation between the multiple linguistic levels of description. With these tools, I proceed to analyzing corpora, investigating the presence of multifractal structures, their relation to linguistic levels of description, and their basic cognitive implications. Finally, the *General Discussion* integrates the results presented here with the known facts found in the linguistic and psycholinguistic literature, and how theories on language processing can be adjusted to account for these findings.

The Constant Entropy Rate Hypothesis

Conditional entropies of linguistic utterances

Before formulating the CER hypothesis in more detail, it is necessary to introduce some notation and concepts from information theory (see Crutchfield & Feldman, 2003 for a detailed review of the relevant information-theoretical concepts used here). A linguistic utterance can be represented as an ordered sequence of n symbols $s_1\, s_2 \ldots s_n$ each drawn from an alphabet Σ containing λ different symbols ($\Sigma = \{\sigma_1,\, \sigma_2, \ldots, \sigma_\lambda\}$). These symbols can be – among many possible choices – either letters, phonemes, morphemes, words, or types of syntactic structures. One can then define the uncertainty about the identity of a symbol to come, provided that the identity of all previously encountered symbols is known. For this, one makes use of the *local conditional (dynamic) entropy* of the continuation of an observed signal

$$\begin{aligned} h^{(n)}(s_1\, s_2 \ldots s_n) &= H(s_{n+1} \,|\, s_1\, s_2 \ldots s_n) = \\ &= -\sum_{i=1}^{\lambda} p\left(\sigma_i \,|\, s_1\, s_2 \ldots s_n\right) \log_\lambda p\left(\sigma_i \,|\, s_1\, s_2 \ldots s_n\right), \end{aligned} \tag{1}$$

where $p\left(\sigma \,|\, s_1\, s_2 \ldots s_n\right)$ denotes the probability that a string $s_1\, s_2 \ldots s_n$ has symbol σ as its immediate continuation. Notice that here, and throughout the rest of this paper, I am using logarithms to the base of the alphabet size λ. Doing this, the conditional entropy measure is bounded to take values between zero and one, and its value does not depend on the cardinality of the alphabet.

More generally, one can consider the average information added by the $(n+1)$-th symbol in the string, that is, the expectation of $h^{(i)}(s_1\, s_2 \ldots s_n)$ over all possible strings of length n, the *conditional (dynamic) entropy*:

$$h^{(n)} = \left\langle h^{(n)}(s_1\, s_2 \ldots s_n) \right\rangle = \sum_{s_1\, s_2 \ldots s_n \in \Sigma^n} p^{(n)}(s_1\, s_2 \ldots s_n)\, h^{(n)}(s_1\, s_2 \ldots s_n), \tag{2}$$

where $p^{(n)}(s_1\, s_2 \ldots s_n)$ denotes the probability of encountering a string $s_1\, s_2 \ldots s_n$ among all the strings of length n. In stationary processes, the average conditional informations must be monotonically decreasing for increasing length, that is, $h^{(n)} \geq h^{(n+1)}$ for all values of n (Crutchfield & Feldman, 2003).

Finally, it is also useful to define the conditional entropy of a context of size k with the additional conditioning on it finishing at position n of the overall utterance

$$\begin{aligned} h^{(n,k)}(s_1\, s_2 \ldots s_k) &= \left\langle h^{(k)}(b_1\, b_2 \ldots b_{n-k}\, s_1\, s_2 \ldots s_k) \right\rangle_{b_1\, b_2 \ldots b_{n-k}} = \\ &= \sum_{b_1\, b_2 \ldots b_{n-k} \in \Sigma^{n-k}} p^{(n)}(b_1\, b_2 \ldots b_{n-k}\, s_1\, s_2 \ldots s_k)\, h^{(n)}(b_1\, b_2 \ldots b_{n-k}\, s_1\, s_2 \ldots s_k). \end{aligned} \tag{3}$$

The CER hypothesis

As formulated in Genzel and Charniak (2002), the CER hypothesis predicts that the average conditional entropies $h^{(i)}$ remain constant for all values of i. More formally, the theory proposes that, for any given linguistic context (*i.e.*, modality, situation, genre, *etc*), there is a constant $0 < C \leq 1$, corresponding to the channel capacity in that situation, such that[1]

$$h^{(i)} = C. \tag{4}$$

Genzel and Charniak (2002) and Aylett and Turk (2004) derive their versions of the CER hypothesis from an assumption of rationality. It is evident that the optimal usage of a channel for information transmission is to transmit the information at a constant rate, approximating the maximum channel capacity in as much as possible. Therefore, if speakers or writers are producing information at the maximum channel capacity, the rate of information transmission should be roughly constant around this maximum capacity, all along the duration of an utterance.

Technically, it is often not feasible to accurately estimate $h^{(i)}$ for large values of i, thus making it difficult to directly test the CER hypothesis using linguistic data. A way around this problem is to decompose de conditional entropy into a component reflecting the immediate past $s_{i-k+1}\, s_{i-k+2} \ldots s_i$, and another component governed by the more distant past $s_1\, s_2 \ldots s_{i-k}$ (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009). The breaking point $i - k$ is can be set as the beginning of the current phrase, sentence, paragraph, *etc*. One can then decompose the conditional entropy as

$$h^{(i)} = h^{(i,k)} - I_{i,k}, \tag{5}$$

where $I_{i,k}$ denotes the mutual information between the distant past $s_{i-k+1}\, s_{i-k+2} \ldots s_i$ and the future s_{i+1}, conditioned on the recent past $s_1\, s_2 \ldots s_{i-k}$.

Although the estimation of $I_{i,k}$ remains as difficult a problem as that of estimating $h^{(i)}$ for large values of i, current techniques from statistical natural language processing provide rather good estimates of $h^{(i,k)}$ when k is kept at reasonably small values $k \ll i$, such as the context of an individual sentence.[2] It is assumed that, as the length of the distant past increases, its informativity about the future grows. This should be reflected in the mutual information between the future and the distant past, conditioned on the recent past, increasing with sequence position. This assumption is somehow supported by the findings that increasing the length of the context considered, one increases the accuracy of predictions in language (Cover & King, 1978; Shannon, 1951; it must be also noted that others have found this context to be limited to 32 letters, beyond which no significant improvement was observed Burton & Licklider, 1955, Moradi, Roberts, & Grzymala-Busse, 1998, and others have claimed that mutual information between words extinguishes beyond five words Huang et al., 1993; Pothos & Juola, 2007). If one accepts this assumption, in order for the $h^{(i)}$ measure to remain constant with increasing i as in Eq. 4, Eq. 5 dictates that the local measure $h^{(i,k)}$ must also increase with i to compensate. Therefore, under the assumption of monotonically increasing $I_{i,k}$, the prediction of increasing $h^{(i,k)}$ provides for a feasible test of the CER hypothesis.

This prediction was tested by studying the evolution of $h^{(i,k)}$ across the sentences in textual corpora (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009). In these studies, the word (defined as anything appearing between two spaces) was considered as

[1] The value of $C = 0$ is not possible, as it would imply a fully deterministic system, which is inconsistent with the known stochastic nature of language.

[2] Although never made explicit, many implicitly make use of an approximation $h^{(i,k)}(s_{i-k+1}\, s_{i-k+2} \ldots s_i) \approx h^{(k)}(s_{i-k+1}\, s_{i-k+2} \ldots s_i)$ (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009) Although these two measures are likely to be very similar, it is important to notice that, strictly speaking, these are different things.

the basic symbol in the sequence, taken from an alphabet including as individual symbols the whole set of words observed in the corpus (with the appropriate Good-Turing corrections to account for existing words that did not appear in the particular corpora studied). They defined the recent past as the elements within the same sentence, and the distant past as the elements in previous sentences. As predicted, they found a monotonically increasing pattern in the local conditional entropy. Sentences appearing later in the corpus had larger average values of $h^{(i,k)}$, which is consistent with the prediction of the CER hypothesis (under the increasing $I_{i,k}$ assumption). The effect is present across languages and text genres (Genzel & Charniak, 2003; Keller, 2004; Qian & Jaeger, 2009), in transcribed spoken dialogues (Vega & Ward, 2009), and significant variation in the strength of the effect is found across different genres (Genzel & Charniak, 2003; Keller, 2004).

The increase in the local complexity of a sentence (without considering the distant past) can arise due to at least two reasons: lexical and syntactic. On the one hand, lexical complexity would be increased by the use of less frequent words. On the other hand, syntactic complexity would be reflected in the use of more elaborate syntactic structures, that is, more intrincate parse trees given a particular grammar. Indeed, it is the case that both of the usage of infrequent words, and the usage of more elaborate parse trees (Genzel & Charniak, 2003; Keller, 2004) seem to increase with sentence position in a text.

Additional support for the CER hypothesis comes from considering the actual length of linguistic units. When a unit is highly predictable from its context (low information content in the sense of Shannon, 1948), it is likely to be expressed in a shorter sequence than in cases when the unit is less expected (high information content). By spreading the information conveyed by a linguistic unit over longer or shorter spaces depending on its total information content, the entropy rate (the average information content per time unit) can be kept constant. Indeed researchers have found this to be the case at multiple levels. Phonemes and syllables occurring in predictable positions receive a shorter and laxer articulation than the same phonemes or syllables occurring in less predictable positions (Aylett & Turk, 2004; Piantadosi et al., 2009). Similarly, words occurring in less predictable positions within a sentence tend to be longer than words in more predictable positions (McDonald & Shillcock, 2001; Piantadosi et al., 2009), and phrases containing predictable elements tend to elide certain function words (Jaeger, 2010; Levy & Jaeger, 2007) or use contracted forms (Frank & Jaeger, 2008) more often than do phrases composed of less predictable elements. Similarly, there tends to be a balance between the number of phonemes and number of syllables in a word, more phonemes per-syllable imply more syllables per word, in what is known as the Menzerath-Altmann Law (Altmann, 1980), and these compensations between the lengths of linguistic units extend across many linguistic scales (Fenk & Fenk-Oczlon, 1993; Fenk-Oczlon & Fenk, 1999, 2002, 2002, 2005, 2007).

Some difficulties of the CER hypothesis

Agreeing that the CER hypothesis presents an interesting direction in the study of language processing and that, as reviewed above, convergent results from different levels of language study are in agreement with it, one must also note some problems with this hypothesis. Here, I introduce three of these problems, in order of increasing importance.

Persistent linear increases in local conditional entropy are not possible. The first – and mildest – problem is mostly a methodological issue. The pattern of increase of the local conditional entropies with sentence position has been observed to be non-linear, with an initial fast increase in the earlier sentences, and a gradual attenuation as discourse develops (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009). However, the pattern is normally studied using a linear regression, and the slopes of this regression are used as indices of the effect (Genzel & Charniak, 2002, 2003; Keller, 2004). This is clearly an inadequate analysis technique for such type of data. The local entropy rate is bounded to take values strictly between zero and one (or between zero and $\log_b \lambda$ if one is using $b \neq \lambda$ as the base of the logarithms). A plain linear growth pattern would have the implication that at some large time the local conditional entropy would exceed its maximum possible value, which is nonsensical unless one is willing to consider languages with infinite vocabularies and unbounded mean word length. The solution for this problem lies in using non-linear regression techniques to characterize the effect, as was also done by Qian and Jaeger (2009). However, rather than using generic polynomial terms, or some type of non-parametric smoother in the regressions (Qian & Jaeger, 2009), an adequate test of the CER hypothesis requires an explicit – parametric – non-linear model of the growth of the local conditional entropies which is theoretically motivated. In this study, I propose models of such type.

The formulation of the CER hypothesis is not consistent. At first sight, the formulation of the CER hypothesis may seem clear and straightforward; the rate at which information is introduced in linguistic utterances is constant along time. However, when one examines in detail the different readings and tests that have been made of the hypothesis, one finds that different authors have actually meant slightly different things when talking of a constant rate. In fact, these differences can in some cases lead to contradictions between them.

For instance, may of the studies have investigated whether the information conveyed by individual words remains constant along discourse (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009). In these cases the basic symbolic unit across which entropy remains constant is the word, that is, the entropy is constant per lexical unit (not per unit of time). On the other hand, some authors have applied a similar principle to sub-lexical levels such as the syllable (Aylett & Turk, 2004). Here, it is found that the same syllable receives a longer realization in more informative contexts than in less informative ones. Therefore, what is being kept constant here is the actual information per unit of time (information in syllable divided by syllable length in ms.). Both types of studies arrive at a similar conclusion, namely that there is an adjustment in unit length to compensate for informational load, which is argued to remain constant. However, the conclusions drawn also have notably contradictory implications. The duration of articulation of words does vary, hence, if two words have different lengths, keeping the information per unit of time constant forcefully means that the total information supplied by the words (rate × duration) is in fact different, which would come in contradiction with the results arguing for a constant rate across words (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009), and in particular for those that make this argument based on actual spoken dialogue data (Vega & Ward, 2009). On the other hand, if one keeps the informational content per word constant, the different lengths of articulation for words of different lengths necessarily imply that the information rate per unit of time will not be constant.

A similar contradiction arises between those studies that formulate the CER hypothesis at the lexical level (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009), and those that argue for its presence at the level of phrases (Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007). As before, the number of words within a phrase is not constant. Therefore, keeping the information content per word constant will lead to variations in the information content per phrase, and *vice-versa*, keeping constant the information conveyed by phrases, will result in variations on the informational content per word.

In sum, a clear formulation of the CER hypothesis requires a specific choice of the units on which the entropy will be constant, and an explicit discussion of the relation between the entropies at different levels of description. In the following sections I provide such a description.

Long-range correlations found in language can contradict the CER hypothesis. The basic idea behind testing the CER hypothesis by looking for a monotonic increase in the local conditional entropy (Genzel & Charniak, 2002, 2003; Keller, 2004; Qian & Jaeger, 2009; Vega & Ward, 2009) is that, as the elements in linguistic utterances are correlated over very long scales (Cover & King, 1978; Shannon, 1951; but see also, Burton & Licklider, 1955; Huang et al., 1993; Moradi et al., 1998; Pothos & Juola, 2007), increasing the context under consideration will provide an improvement in the predictive power, and thus the expected increase in $I_{i,k}$. Notice that if these correlations disappear for some (possibly large) value $k = \tau$, the increase in $I_{i,k}$ would also dissappear for all $k > \tau$, and therefore there would be no reason to expect an increase in the local conditional entropy $h^{i,k}$ for these values. In sum, the test devised by Genzel and Charniak (2002) implicitly relies on the presence of possibly infinite correlation lengths τ. That is to say, any two symbols in a linguistic stream should be positively correlated, no matter how large the number of intervening symbols separating them. Such long-range correlations (LRC), alternatively termed $1/f^{\alpha}$ noise or long memory, are the hallmark of stochastic fractals (Mandelbrot & Van Ness, 1968).

Indeed, linguistic utterances exhibit fractal style LRC, both when considering actual speech signals (Kello et al., 2008; Lardner et al., 1992; Maragos & Potamianos, 1999; Pickover & Khorasani, 1986; Sabanal & Nakagawa, 1996; Voss & Clarke, 1975), and when studying texts as symbolic sequences (Amit et al., 1994; Dębowski, 2006; Ebeling, 1997; Ebeling & Frömmel, 1998; Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Gillet & Ausloos, 2008; Hilberg, 1990; Kokol & Podgorelec, 2000; Li, 1989; Pavlov et al., 2001; Schenkel et al., 1993). This could lead us to feel safe on the plausibility of the CER hypothesis. However, a crucial implication of systems with LRC is that, as long as they are ergodic and stationary (which are two assumptions that are implicit in all tests of the CER) their entropy rate is strictly decreasing with increasing context length, with the decrease following a power-law structure (Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Hilberg, 1990). For ergodic stationary sequences, the LRC properties of linguistic utterances are in strong contradiction with the CER hypothesis: the CER hypothesis *cannot possibly hold* in the presence of LRC. Therefore, the only way of reconciling the CER hypothesis with the LRC property, is to assume that linguistic sequences are either not ergodic, or are non-stationary. Below, I provide converging evidence indicating that language sequences are indeed fundamentally non-stationary in non-trivial ways.

Another way to regard this contradiction between LRC and the CER hypothesis is to notice that an ‘optimal’ use of the channel would in fact use a coding that maximally exploits the long-correlations, and by doing so removes them from the sequence. In other words, that would mean that the human language system would work as an online data compressor/decompressor; compression algorithms enable a maximal use of a channel’s capacity, and they do so by exploiting and removing the correlations in the data. Instead, if something characterizes human language is the large amount of redundancy that is uses. This redundancy enables to achieve a robust – rather than just efficient – communication.

Theoretical Concepts & Analyses

In this section I introduce further measures from Information Theory and Signal Processing that are useful for characterizing the LRC properties of a symbolic dynamical system. I introduce the concepts of block entropy, entropy of the source, lagged transinformations, the Fourier spectrum of a symbolic sequence, its Lempel-Ziv complexity, together with their relation to the LRC property and, importantly, the implications on the possible validity of the CER hypothesis and language processing that can be derived using these techniques.

Block entropies & dynamical entropy of the source

A useful quantity in the study of symbolic complex systems are the *block entropies* H_n at different lengths n, that is, the uncertainty on the identity of a string of length n in the language,

$$H_n = - \sum_{b_1\, b_2 \ldots b_n\, \in \Sigma^n} p^{(n)}(b_1\, b_2 \ldots b_n) \log_\lambda p^{(n)}(b_1\, b_2 \ldots b_n), \tag{6}$$

and the block entropy at length one is given by

$$H_1 = - \sum_{j=1}^{\lambda} p(\sigma_j) \log_\lambda p(\sigma_j). \tag{7}$$

Notice that, by definition, the conditional entropy can be expressed in terms of the block entropies:

$$h^{(n)} = H_{n+1} - H_n. \tag{8}$$

From the block entropies one can also define the *entropy rates*

$$H^{(n)} = \frac{H_n}{n} \tag{9}$$

Taking the conditional entropies to their infinite length limit, one obtains the *dynamical entropy of the source* (h; the Kolmogorov-Sinai entropy; Khinchin, 1957), which is a fundamental measure for the characterization of a dynamical system:

$$\begin{aligned} h \quad &= \lim_{n\to\infty} \left[h^{(n)}\right] \quad = \lim_{n\to\infty} \left[H_{n+1} - H_n\right] = && \text{(10a)} \\ &= \lim_{n\to\infty} \left[H^{(n)}\right] \quad = \lim_{n\to\infty} \left[\frac{H_n}{n}\right]. && \text{(10b)} \end{aligned}$$

This quantity corresponds to the asymptotic rate of entropy production for infinitely long strings.

Estimating the block entropies is not easy from actual texts. The number of different strings of length n found in the text (the n-*words*) grows very fast with n. This number is approximately λ^{H_n} (Ebeling & Nicolis, 1992), which can be of astronomical proportions already for moderate values of n, with the situation becoming worse in more chaotic systems, for which H_n grows linearly with n. For a typical written alphabet, it becomes computationally impractical to keep track of the frequencies of that many possibilities for values of n larger than around 40. Furthermore, due to finite sample effects, this method breaks down from $n \approx \log_\lambda N$ onwards, which with the typical alphabet size (*i.e.,* $\lambda = 29$ in the empirical studies below) is of the order of $n = 4$ even for very large texts (Schürmann & Grassberger, 1996). Some degree of correction for these finite sample effects can be achieved by using alternative methods based on counting the number of distinct words of a given length that appear in the text (Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995), and by using the estimates obtained from multiple texts of different lengths (Grassberger, 1988; Herzel, 1988; Herzel, Schmitt, & Ebeling, 1994; Li, 1990), but the estimations remain very noisy even for moderate n. Below, I introduce alternative methods to investigate the scaling at longer scales.

Scaling in block entropies

In general, the scaling of the block entropy of a stochastic symbolic system can be modeled as a function of the string length n by the expression (Crutchfield & Feldman, 2003; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Grassberger, 1988):

$$H_n = h\,n + gn^{\mu_1}\,(\log n)^{\mu_2} + e, \tag{11}$$

with either $0 \leq \mu_1 < 1$, or $\mu_1 = 1$ and $\mu_2 < 0$.

Using these expressions, the LRC property can be characterized by the asymptotic behavior of the entropies for large n. The coefficient h is the dynamical entropy of the source. This coefficient is positive ($h > 0$) for any stochastic or chaotic system. Markovian systems with a memory limited to m steps are characterized by $e > 0, g = 0$, so that their uncertainty decreases during the first m steps, to then settle in $h^{(n+m)} = h^{(n)} = h$. The limiting case of these are fully chaotic, uncorrelated, systems, represented by the case when $g = 0$ and $e = 0$. In such systems the block entropy increases at a constant rate $h > 0$, that is, $h^{(n)} = h$. Finally, a deterministic system with a period m is characterized by $H_{n+m} = H_n = e$.

The more interesting case are those string that are neither Markovian or periodic, nor fully chaotic, what is sometimes referred to as the borderline between order an chaos. Among these, of particular interest are the cases:

(a) Logarithmic tails:

$$H_n = h\,n + (\log n)^{\mu_2} + e, \quad \mu_2 < 0 \tag{12}$$

(b) Power-law tails:

$$H_n = h\,n + gn^{\mu_1} + e, \quad 0 < \mu_1 < 1 \tag{13}$$

It has been argued that such intermediate situations are prototypical of information carrying sequences. In particular, texts written in English and German have been found to exhibit the situation (b), with a scaling exponent $\mu_1 = 1/2$, which would be indicative of LRC (Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Hilberg, 1990).

The above discussion implies that that the conditional entropy remains constant along a text, as would be predicted by the CER hypothesis, is only possible in the cases when the system is fully chaotic ($h^{(n)} = h,\, \forall n$) or Markovian with a memory length m ($h^{(n)} = h,\, \forall n > m$). However, as it has been noted by many – and I will corroborate below – linguistic sequences are not Markovian, but rather exhibit LRC.

As I will discuss below, it is quite possible that the scaling indices for linguistic sequences change at different scales, as is characteristic of multifractal systems. It is therefore important to assess the scaling properties at multiple values of n, both small and large. In order to achieve more accurate descriptions of the scaling regimes for moderate and large values of n, I now introduce the lagged transinformations and the symbolic Fourier spectra.

Lagged transinformations

The LRC property is defined by the never-extinguishing correlations between lagged pairs of items in a sequence. In the case of symbolic sequences, it is natural to consider, instead of the correlation, the mutual information between pairs of symbols at different distances (Ebeling & Frömmel, 1998; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Li, 1989, 1990; Li & Kaneko, 1992). These are the *lagged transinformations* at lag $n \geq 1$:

$$I_n = \sum_{i=1}^{\lambda} \sum_{j=1}^{\lambda} p^{(n+1)}(\sigma_i, \sigma_j) \log_\lambda \frac{p^{(n+1)}(\sigma_i, \sigma_j)}{p(\sigma_i)\, p(\sigma_j)}, \tag{14}$$

where $p^{(n+1)}(\sigma_i, \sigma_j)$ denotes the probability of encountering a string of length $n+1$ with σ_i as its first symbol and σ_j as its last one:

$$p^{(n+1)}(\sigma_i, \sigma_j) = \sum_{b_1 \dots b_{n-1} \in \Sigma^{n-1}} p^{(n+1)}(\sigma_i\, b_1 \dots b_{n-1}\, \sigma_j),$$

and $p^{(2)}(\sigma_i, \sigma_j) = p^{(2)}(\sigma_i\, \sigma_j)$.

In systems that exhibit LRC, the lagged transinformation should also scale as a power-law of the lag,

$$I_n \propto n^{-\gamma},\ \gamma \geq 0, \tag{15}$$

so that no matter how far two elements are in a sequence, they will contain information about each other. Therefore, this measure can also be used to assess the LRC property in language sequences, with the advantage that its computation is much simpler than that of the block entropies, especially so for large values of n. Indeed, it has been found by many that language sequences exhibit LRC-like scaling of the transinformations (Li, 1989; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Ebeling & Frömmel, 1998), but some have recently claimed that the mutual information is not significant beyond lags of five words (Huang et al., 1993; Pothos & Juola, 2007).

The transinformations are also subject to finite sample effects, estimates obtained from a finite text are larger than the actual value that would be obtained for a hypothetical infinite-length text. In particular, it is known that the mean overestimation of I_n due to the finite length L of the text from which the estimates were obtained is given by (Herzel, 1988; Herzel et al., 1994; Li, 1990)

$$\delta I_n(L) = \left\langle \hat{I}_n(L) - I_n \right\rangle \approx \frac{\lambda^2 - 2\lambda}{2L \ln \lambda}, \tag{16}$$

where $\hat{I}_n(L)$ denotes the estimate of the n-lagged transinformation obtained from a text of length L, λ refers to the cardinality of the alphabet, and the operator $\langle \cdot \rangle$ denotes the expected value (*i.e.,* the average) of a function.

Notice that Eq. 16 provides a straightforward method for correcting for the finite sample overestimation. As L grows to infinity, $\delta I_n(L)$ goes to zero as a linear function of $1/L$, and the estimate of I_n becomes more accurate. Therefore, if one performs a linear regression between $1/L$ and the estimate $\hat{I}_n(L)$ using multiple lengths, extrapolating to the intercept of the regression line with the vertical axis ($1/L = 0$) estimates the value of I_n when $L \to \infty$, providing also the corresponding standard errors of the estimate.

The correction above enables to estimate the lagged transinformations for a large range of values. However, the standard errors of the intercept estimates in the regressions still limit the method so that the corrected transinformation estimates become inaccurate at very long lags, where it is still not possible to distinguish them from random variation. In addition, although the presence of LRCs is clearly indicated by having a scaling with $\gamma > 0$, for sequences other than binary sequences ($\lambda > 2$), it is difficult to achieve a detailed interpretation of the value of γ (Li, 1990). For these reasons, I introduce below the symbolic Fourier spectra, which are more accurate in detecting the scaling at the very low frequencies (*i.e.*, very long distances), and enables a detailed interpretation of the scaling exponents.

Symbolic Fourier spectra

Perhaps the most common way of studying the LRC property in time series is by looking for a power-law scaling relation in the Fourier spectrum of the series. However, in the case of symbolic sequences, the definition of the Fourier spectrum is less straightforward than it is in numerical sequences. Several techniques have been developed for computing the spectrum of a symbolic sequence (Afreixo, Ferreira, & Santos, 2004; Li & Kaneko, 1992; Silverman & Linsker, 1986; Stoffer, Tyler, & McDougall, 1993; Voss, 1992; Wang & Johnson, 2002). Here, I will consider the approach using indicator sequences that was first introduced by Voss (1992).

A sequence of symbols $\mathbf{s} = s_1\, s_2 \ldots s_n$ with each symbol sampled from an alphabet $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_\lambda\}$ can be mapped into λ binary indicator sequences $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(\lambda)}$, such that the i-th element of the j-th indicator sequence is given by:

$$u_i^{(j)} = f_j(s_i), \qquad f_j(x) = \begin{cases} 1, & \text{if } x = \sigma_j \\ 0, & \text{if } x \neq \sigma_j \end{cases}, \tag{17}$$

where s_i is the i-th element of the original sequence $\mathbf{s}$ and σ_j is the j-th symbol in the alphabet Σ. In this way, the original symbolic sequence is converted into a trajectory in a λ-dimensional binary space. In short, the original sequence is converted into a $n \times \lambda$ matrix, the rows of which correspond to each position in the sequence, and the columns $(\mathbf{u}^{(j)})$ correspond to each of the symbols in the alphabet. The element in position (i, j) in this matrix is one if the i-th symbol of the original sequence is σ_j, and zero otherwise.

Now, for each of the binary indicator sequences, the corresponding Discrete Fourier Transforms (DFT) can be computed using the normal methods (usually the FFT algorithm), so that $\mathbf{U}^{(i)}$ is the DFT of the indicator sequence $\mathbf{u}^{(i)}$. From here, the i-th component of Fourier spectrum of the symbolic sequence is computed as

$$S_i = \sum_{j=1}^{\lambda} \left\| U_i^{(j)} \right\|^2, \tag{18}$$

where $\|\cdot\|$ denotes the modulus of a complex number, and $U_i^{(j)}$ is the i-th element of the DFT of the j-th indicator sequence. In short, the spectrum of the symbolic sequence is given by the sum of the spectra of the indicator sequences corresponding to each symbol in the alphabet.

As was advanced above, series exhibiting LRC (*i.e.*, symbolic $1/f$ noise; Li, 1990) exhibit power spectra that scale as a power-law of the frequency. In terms of the symbolic Fourier spectrum this can be expressed as,

$$S_f \propto f^{-\beta}, \tag{19}$$

where f are the elements (frequencies) of the spectrum. As in the case of usual $1/f$ noise, the absence of LRC is signaled by observing a scaling of the spectrum with $\beta = 0$. Therefore, if one observes $\beta \neq 0$, the CER hypothesis cannot hold, as the system is neither fully chaotic nor Makovian, which as I discussed above, are the only situations in which one can expect the conditional entropies to keep constant rather than decrease with time. In human writings, values of β different from zero have indeed been reported by several authors (Li, 1989; Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995).

Lempel-Ziv complexity & decomposition of information sources

The techniques above enable to assess the scaling of linguistic sequence, but it will be also useful to have an estimate of their limit entropy of the source. With this goal, I know consider a completely different approach for estimating the entropy of the source. The definition for the entropy of the source and the associated measures used above rely on Shannon's approach to complexity. These are based in estimating global features of the system's dynamics, that is, instead of relying on the properties of a single sequence, the ergodic assumption is used to ensure that the probabilities found in one very long sequence will mimic the probabilities of an ensemble of sequences. In contrast, the Lempel-Ziv and Ziv-Lempel data compression algorithms (LZ76, ZL78; Lempel & Ziv, 1976; Ziv & Lempel, 1977, 1978) offer a different route for estimating h, explicitly relying only on the properties of a single sequence. The compressibility of a sequence is the size of a compressed version relative to the original sequence. For stationary ergodic sequences, it is known that, on the limit of infinite sequence length, the compressibility of a sequence using these algorithms converges to the entropy of the source measured in $\log_\lambda$ units (Ziv & Lempel, 1978).

In the LZ76 and the ZL78 algorithms, a given sequence of n symbols $s_1 s_2 \ldots s_n$ is parsed in a sequence of $\mathcal{N}$ words $w_1 \cdot w_2 \cdot \ldots \cdot w_{\mathcal{N}}$. In both cases the first word is $w_1 = s_1$, however, there is a crucial difference in the way the remaining words are chosen. In the original LZ76 algorithm and w_{i+1} is the shortest word (starting immediately after w_i) that has not been encountered in the string before. In the simplified ZL78 version, it suffices that is different from all previously used words, but it may have previously occurred in the string. As an illustration of the difference in codings, according to the simpler ZL78 algorithm, the binary sequence $10011011100101000100\ldots$ would be parsed into the sequence of words $1 \cdot 0 \cdot 01 \cdot 10 \cdot 11 \cdot 100 \cdot 101 \cdot 00 \cdot 010 \cdot 0 \ldots$ On the other hand, following the LZ76 algorithm, the parsing would have been $1 \cdot 0 \cdot 01 \cdot 101 \cdot 1100 \cdot 1011 \cdot 00100 \cdot \ldots$ The difference between the two parses becomes evident from the fourth word in each. Whereas the ZL78 uses 10, as this is the shortest non previously used word, the original LZ76 uses 101, as 10 has already been observed in the string. It is not difficult to see that the LZ76 algorithm will result in longer words, and hence a lower value of $\mathcal{N}$.

The rationale behind these decompositions is intuitive. The algorithms look for a recoding of the sequence into a new set of symbols (the words) all of which have exactly the same probability of occurrence. In this way, it looks for the length of an equivalent sequence that has a maximal entropy rate. Using either algorithm, the *Lempel-Ziv complexity* is then given by the expression

$$L_n = \frac{\mathcal{N}(1 + \log_\lambda \mathcal{N})}{n}, \tag{20}$$

where, as usual, λ denotes the cardinality of the alphabet. Crucially, for very long sequences, the Lempel-Ziv complexity converges to the entropy of the source (Ziv & Lempel, 1978):

$$L_\infty = \lim_{n\to\infty} L_n = \lim_{n\to\infty} \frac{\mathcal{N}(1 + \log_\lambda \mathcal{N})}{n} = h. \tag{21}$$

Although the approach of L_n to its asymptotic value of h is relatively fast, estimates of h based directly on L_n for some large n are still subject to finite sample effects. To correct for these, one can use the empirical approximation (Schürmann & Grassberger, 1996)

$$L_n \approx h + a\, n^{-b} \ln n, \quad b > 0, \tag{22}$$

where a and b are parameters that can be fitted from data. To do this, give a sequence of length n, one computes L_{n_i} for a range of k subsequences from the original one, with increasing lengths $n_1 < n_2 < \ldots < n_k = n$, and a least-squares non-linear regression on these values estimates the parameters a and b, together with the asymptotic value of h. This approximation was proposed by Schürmann and Grassberger (1996) based on empirical observation, but without a theoretical motivation. I have confirmed that it provides a very accurate description of the evolution of the Lempel-Ziv complexity, with explained variance values of above 98%.

An important remark must be made here. Some researchers have used the simpler ZL78 algorithm (Ziv & Lempel, 1977, 1978) variant of the compression algorithm for estimating the entropy rate of texts (Juola, 1998; Juola, Bailey, & Pothos, 1998; Juola, 2007). However, it has been shown that this variant of the algorithm is very inaccurate in estimating h (Amigó, Szczpański, Wajnryb, & Sánchez-Vives, 2004), most markedly for sequences that exhibit LRC, whereas the original LZ76 remains very accurate even in these situations (Lesne, Blanc, & Pezard, 2009). In those cases where the complexity was estimated using the widely available `gzip` compression software (Juola, 1998, 2007), the inadequacy of the ZL78 algorithm is compounded with the implementational limitations of `gzip`, which make it unsuitable for entropy estimation.[3]

The Lempel-Ziv technique provides an estimate of the entropy of the source. Entropy is actually the opposite of information; that is to say, information in a signal is measured in entropy decreases that reflect its departure from randomness. Therefore, the information in a linguistic signal is given by its redundancy $I = 1 - h$. One can think of a linguistic sequence as the superposition between multiple sources of information: lexical, morphological, syntactic, pragmatic, *etc.* If one further assumes that these sources of information are relatively independent, one can take them to be additive (Juola, 1998, 2007). This view, that I will exploit in the analysis section, enables the decomposition of the source entropy, so that each of the levels contributes some degree of information. Given a total maximum entropy of one, each of the levels reduces this uncertatity

$$I_{\text{total}} = 1 - h = I_{\text{lexical}} + I_{\text{syntax}} + I_{\text{pragmatic}} + \ldots$$

Now, if one degrades the signal removing only the information from one the levels, say for instance the syntactic structure, one should observe

$$I_{\text{degraded}} = 1 - h_{\text{degraded}} = I_{\text{lexical}} + I_{\text{pragmatic}} + \ldots$$

so that one can estimate the amount of information provided by the syntax alone (or any other level that has been disturbed) as

$$I_{\text{syntax}} = I_{\text{total}} - I_{\text{degraded}} = h_{\text{degraded}} - h. \tag{23}$$

As I will show below, this can provide very valuable information on the structure of texts.

[3] A good overview of these problems, with an annotated bibliography, is provided by C.R. Shalizi in `http://cscs.umich.edu/~crshalizi/notebooks/cep-gzip.html` and `http://www.cscs.umich.edu/~crshalizi/notabene/entropy-estimation.html`.

Extension from characters to words

In the empirical section, as was done previously (Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Ebeling, 1997; Ebeling & Frömmel, 1998; Hilberg, 1990; Kokol & Podgorelec, 2000; Li, 1989; Schenkel et al., 1993; Pavlov et al., 2001), I will study the scaling properties of texts using the characters as the basic symbols in the alphabet. It is however important to note that the conclusions reached on characters can be extrapolated to any other arbitrary basic units. Here I illustrate how to draw conclusions on the behavior of the system at the word level from the results on the character level. Evidently, an identical approach can be used to take conclusions on other linguistic levels of study, either more fine-grained, as in morphemes, or more coarse-grained as in the case of phrases or sentences.

If one has used characters as the basic unit of study, one can use the block entropies to measure the average information contributed by a word of length l starting at position $n+1$ of the sequence,

$$h_w^{(n,l)} = H_{n+l} - H_n, \tag{24}$$

defining also $H_0 = 0$. This enables to define the average contribution of a word at position n, that is, the conditional entropy of the text considered as a sequence of words instead of as a sequence of letters,

$$h_w^{(n)} = \left\langle h_w^{(n,l)} \right\rangle_l = \sum_{l=1}^{\infty} p(l, n+1)\, h_w^{(n,l)}, \tag{25}$$

where $p(l, n+1)$ denotes the probability of finding a word of length l starting at position $n+1$ in the text.

Taking the character as the basic unit, the discussion above shows that the LRC property implies that it is not possible that the conditional entropy of the text remains constant, as is proposed by the CER hypothesis. It is, however, still possible that the conditional entropy keeps constant at coarser grained level of description, such as words. This could happen if a suitable modulation of the probabilities $p(l, n+1)$ in Eq. 25 balances out the decrease in the per letter conditional entropy. Notice here that this would already be a different theory than that advocated by the proponents of the CER hypothesis. Whereas the CER argues for a constant entropy rate across levels, here we would be presenting a weaker form of the theory, where the CER would only hold at one particular level of description (*e.g.,* words), hence avoiding the second problem of the CER hypothesis mentioned in the previous section. I will refer to this weaker form of the theory as the *relaxed CER* hypothesis (rCER). It is important to realize that such adjustments in word length imply that the sequences will be non-stationary. At the level of characters, increasing the average word length comes with decreasing the frequency of spaces and punctuation marks. At the level of words, this implies that the probabilities of longer words relative to the shorter ones change throughout the texts. In both levels the probabilities of the basic symbols become dependent on the position in the text. The non-stationarity of the process is crucial. For stationary processes, the conditional entropy is bound to be monotonically decreasing, $h^{(n)} \geq h^{(n+1)} \geq h$, for all values of n (for a demonstration, see *Lemma 1* in Crutchfield and Feldman (2003)). As we will see below, the rCER hypothesis may require the presence of increases in $h^{(n)}$ with n, and this is only possible in non-stationary processes.

Assuming the power law scaling of the conditional entropy that was previously found (Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995;

Hilberg, 1990), combining Eq. 13 with Eqs. 24, 25, one obtains that, for $n \gg 0$,

$$\begin{aligned} h_w^{(n)} &= \sum_{l=1}^{\infty} \Big[p(l, n+1) \big(l\, h + g\, [(n+l)^{\mu_1} - n^{\mu_1}] \big) \Big] = \\ &= h \sum_{l=1}^{\infty} l\, p(l, n+1) + g \sum_{l=1}^{\infty} \Big(p(l, n+1) \big[(n+l)^{\mu_1} - n^{\mu_1} \big] \Big) = \\ &= h \big\langle l(n+1) \big\rangle + g \sum_{l=1}^{\infty} \Big(p(l, n+1) \big[(n+l)^{\mu_1} - n^{\mu_1} \big] \Big), \end{aligned} \tag{26}$$

where h is the per letter entropy of the source, $g > 0$, $0 < \mu_1 < 1$, and $\big\langle l(n+1) \big\rangle$ is the mean length of the words starting at position $n + 1$ in the text. The second term in this equation can be simplified using the approximation,

$$\sum_{l=1}^{\infty} p(l, n+1)(n+l)^{\mu_1} \approx \Big(n + \big\langle l(n+1) \big\rangle \Big)^{\mu_1}, \tag{27}$$

leaving Eq. 26 as[4]

$$h_w^{(n)} \approx h \big\langle l(n+1) \big\rangle + g \left[\Big(n + \big\langle l(n+1) \big\rangle \Big)^{\mu_1} - n^{\mu_1} \right]. \tag{28}$$

Therefore, the evolution of the mean word length along the positions in a text will have a crucial effect on the evolution of the word-based conditional entropy. If the mean word length remains constant along the text then, for large n, the second term of Eq. 28 scales as a power law with exponent $\mu_1 - 1$, such that

$$h_w^{(n)} \approx h \,\langle l \rangle + g_2 n^{\mu_1 - 1}, \tag{29}$$

where $\langle l \rangle$ is the mean word length across all text positions, and the value of $g_2 > 0$ depends on $\langle l \rangle$. In this case, the LRC at the letter level imply an identical scaling at the word level, the conditional entropy would also be monotonically decreasing with n, and the rCER hypothesis would remain invalid. Similarly, if the mean word length decreased monotonically along the text positions, the decrease in the conditional entropy would become even more marked, as it would now affect both terms in Eq. 28, and in the long n limit – as the minimum possible word length is approached – would in any case reduce to a version of Eq. 29 with a perhaps slightly different value of μ_1.

The only option left for balancing the decreasing conditional entropy is therefore to have a monotonic increase in the average word frequency. The growth of the mean word length must be sublinear with text position, as a linear increase would result in unrealistically long words all too quickly. Hence, the fastest rate of increase that one could imagine is a power-law of the type

$$\big\langle l(n) \big\rangle = l_{\min} n^{\eta}, \qquad l_{\min} > 1, \quad 0 < \eta < 1, \tag{30}$$

where $l_{\min}$ reflects the minimum possible mean word length, and η measures the speed of increase of the word length. Under this scaling, for large n, Eq. 28 reduces to

$$h_w^{(n)} \approx h\, l_{\min} n^{\eta} + g_2 n^{\mu_1 - 1 + \eta}. \tag{31}$$

[4] The approximation in Eq. 27 represents an overestimation with an error of approximately $\varepsilon(n) = k\, n^{\mu_1 - 2}$ where k is a constant $0 < k < 1$. For large n, ignoring this term is harmless, as the remaining terms will dominate Eq. 28.

The constraints that the values of η and μ_1 fall in the range between zero and one have the implication that, for large n, the first term will dominate this expression, resulting in an unbounded growth of the conditional entropy, which would also contradict the rCER hypothesis. Furthermore, notice that in such a model, the mean word length is actually unbounded, which has the side implication of dealing with a language with an infinite vocabulary, which does not appear to be realistic.

If one further constrains human languages to have finite vocabularies, then one needs to establish a scaling in which the word length has a bounded growth, that is, although the average length grows monotonically along position, this growth asymptotes to some maximum mean word length $l_{\max}$. The speed at which this maximum word length is approach will determine the situation with the entropy of the source. If this approach is very fast, from some position p onwards, the asymptote has already been reached, and thus, for large n, the system behaves as the constant mean word length situation of Eq. 29, thus again making the rCER hypothesis impossible. On the other hand, a very slow approach to asymptote could eventually approximate the balance that is required for the rCER hypothesis. One can model these two situations as another type of power-law growth:

$$\begin{aligned} \langle l(n) \rangle &= l_{\min} + \Delta_l (1 - n^{-\eta}), \qquad & l_{\min} \geq 1,\, \eta > 0, \\ \Delta_l &= l_{\max} - l_{\min}, \qquad & l_{\max} > l_{\min} \end{aligned} \tag{32}$$

where $l_{\min}$ and $l_{\max}$ reflect the minimum and the maximum possible mean word lengths, and η measures the speed of growth of the mean word length along a text. In this situation, Eq. 28 becomes

$$h_w^{(n)} \approx h\, l_{\max} - h\, \Delta_l (n+1)^{-\eta} + g \left(\left[n + l_{\max} - \Delta_l (n+1)^{-\eta} \right]^{\mu_1} - n^{\mu_1} \right). \tag{33}$$

The first term corresponds to the asymptotic value of the conditional entropy (per word) at infinite time (when the mean word length reaches its maximal value of $l_{\max}$), it is thus the value of the entropy of the source in terms of words h_w. If $\eta > 1$, the approach of the word length to the theoretical maximum is very fast, leading at large n to a situation of basically constant word length, and thus decreasing h_w. In the case of a slower approach – with $\eta < 1$ – Eq. 33 gives rise to a relatively fast decrease of $h_w^{(n)}$ at the small values of n. At very large values of n, there is a global minimum from where it follows a very slow increase towards the asymtotic value of h_w, the speed of which decreases as a power law of the type $n^{-\eta}$. Therefore, even in this limiting case of slow asymptotic increase of the maximum mean word length, the per-word conditional entropy will always be either decreasing or increasing along a text, but never constant. In short, the LRC property found at the letter level excludes all versions of the rCER at higher levels as well. In the empirical section I will illustrate the evolution of $h_w^{(n)}$ with text position using empirical estimates of the parameters h, g, $l_{\min}$, $l_{\max}$, μ_1, and η.

Multifractality

The discussion above has advanced the possibility that linguistic sequences exhibit the LRC property. I have described the LRC property in terms of the scaling exponents μ_1, γ, and β, with the implicit assumption that these take unique values for a given system. In systems with the LRC attribute, a single value for each of the exponents – ultimately reflected in a unique value for a fractal dimension – is prototypical of a *monofractal* system. However, there exist also systems with LRC for which the value of the scaling exponents is not constant,but rather changes in different parts of the system, these are the so-called *multifractal* systems (Stanley & Meakin, 1988). In such systems, rather than a unique value for a fractal dimension, their description requires a singularity spectrum that details the local values of the fractal dimension around a particular point. Symbolic sequences exhibiting multifractal structure are characterized by being *non-stationary* (*i.e.*, their statistical properties change with time) and *inhomogeneous* (*i.e.*, their composition in terms of symbols at different scales changes with time; Ivanov et al., 1999).

In most cases, the singularity spectra cannot be directly observed. Instead one can observe multiscaling spectra, that is, that the values of the scaling exponents change depending of the scale at which the system is being considered. This phenomenon is sometimes termed statistical intermittency, and it denotes that the statistical properties of the system can change dramatically depending on the scale. The statistical intermittency can be observed as a piece-wise linear composition of the different spectra of the signal.

Samples of human language can be expected to be non-stationary and inhomogeneous (Schenkel et al., 1993; Schürmann & Grassberger, 1996; Pavlov et al., 2001), as the topic, lexical composition, style, *etc*., can often change along discourse. Also, linguistic sequences are the result of the superposition of multiple levels of structure (*i.e.,* phonological, lexical, morphological, syntactic, pragmatic, ...) operating at different scales and not necessarily sharing the same properties. Indeed, multiple sources of evidence suggest that linguistic sequences show multifractal structure (Lardner et al., 1992; Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Pavlov et al., 2001; Gillet & Ausloos, 2008).

Corpus Analyses

In this section I will investigate in more detail the hypotheses stated above. Here, I will consider texts as plain sequences of characters. By studying the lagged transinformations, the symbolic Fourier spectra, and the Lempel-Ziv complexity, on textual corpora, I show here that:

1. Linguistic sequences exhibit LRCs, which are incompatible with the CER hypothesis.
2. The structure in texts and dialogues is actually multifractal, with a certain degree of language independency.
3. The different scales correspond to traditional levels of linguistic description, namely lexical, syntactic, and pragmatic levels.
4. An important consequence of the multifractal structure is that texts are markedly non-stationarity. This is the result the long-scale structuring of discourse, and possibly of online cognitive adjustments attempting to optimize channel use.

Corpora & preprocessing

For the main analyses, I made use of the *Europarl Corpus* (Koehn, 2005), version 5. This corpus contains transcripts of sessions of the European Parliament between 1998 and 2009. The sessions are transcribed (some partly) in eleven of the official languages of the E.U.

I selected the English, Finnish, and Greek transcripts of the sessions in the year 2000 (these languages were chosen as representing a maximum of linguistic diversity among the eleven available). Excluding one extremely short session, this contained 69 sessions of the Parliament. Each session is contained in a separate file and was processed individually. The results presented below correspond to averages across the 69 sessions. This is particularly useful, as each session can be considered as a separate independent instance of the strings produced by the same dynamical system. In this sense, the set of sessions can be considered an ensemble of ‘trajectories’ of the system.

In order to separate the contribution of different levels of linguistic structure, in a spirit similar to that of previous studies (Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Juola, 1998, 2007), four versions of the English corpus are considered:

- *Original*: the corpus files (cleaned of annotations) as they are provided.
- *Sentence shuffled*: the order of the sentences in each of the sessions has been randomized. The English sentence splitting tool provided with the corpus was used for identifying sentence boundaries, that removes the ambiguity of the stop sign by analyzing the different contexts in which it can appear (i.e., abbreviations, numbers, *etc.*).
- *Word shuffled*: the order of the words *within each sentence* is randomized. Words were defined as individual tokens provided by the English tokenizer tool that is provided with the corpus.
- *Letter shuffled*: the order of letters within a session was randomized (across sentences and words).

This choice of randomizations enables the separation of three levels of linguistic structure (on top of the information provided by the plain distribution of letter frequencies):

- *Lexical/morphological* structure is preserved in all but the letter shuffled versions of the corpora.
- *Syntactic* structure is preserved in the original and sentence shuffled versions of the corpus (all sentences are instances of well-formed syntactic structures of the language), but lost in the letter shuffled and word shuffled versions.
- *Pragmatic* structure is preserved in the original and word shuffled versions of the corpora (the words are shuffled *within* sentences, but the topical order of the sentences is preserved), and is lost in the letter shuffled and sentence shuffled versions (although the sentences are well-formed, their topical order is destroyed by the shuffling).

In addition, the letter shuffled version of the corpus provides a baseline in which no LRC are present, but that has the same statistical distribution of letters as has the original.

In order to rationalize the number of indicator sequences necessary for the computation of the symbolic Fourier spectra, prior to the spectral estimation, the texts were simplified by converting all letters to lower case, substituting all punctuation marks by a symbol ‘P’, all spacing marks (carriage returns were plainly removed) by a symbol ‘S’, and all digits by a symbol ‘D’. All symbols not belonging to any of those categories were not considered in the Fourier analyses (but the transinformations were computed on the raw text). This reduced the alphabet used in the symbolic spectra to 29 symbols ($\lambda = 29$), that is, 29 indicator sequences.

In addition to the *Europarl* corpus, to provide a indication of the variation of the scaling behavior across styles, I also analyze two samples of literary works downloaded from the *Project Gutenberg*’s website[5]. The first sample contains English language novels, including “Alice’s Adventures in Wonderland” by L. Carroll, “The Hound of Baskervilles” by A. Conan-Doyle, “The Heart of Darkness” by J. Conrad, “Emma” by J. Austen, “White Fang” by J. London, “Ivanhoe” by W. Scott, “Moby Dick” by H. Melville, “Peter Pan” by J. M. Barrie, “Secret Adversary” by A. Christie, “The Treasure Island” by R. L. Stevenson, “A Tale of Two Cities” by C. Dickens, “Uncle Tom’s Cabin” by H. B. Stowe, and “Wuthering Heights” by E. Brontë. The second is a sample of literary works by the Finnish author J. Aho, including: “*Juha*”, “*Katajainen Kansani*”, “*Lohilastuja ja Kalakaskuja*”, “*Muistatko*”, “*Muistelmia ja Matkakuvia*”, “*Panu*”, “*Papin Rouva*”, “*Papin Tytär*”, “*Rauhan Erakko*”, “*Rautatie*”, and “*Yksin*”.

[5] `http://www.gutenberg.org/wiki/Main_Page`

Methods

Lagged transinformations. The single symbol probabilities $p(\sigma)$ were estimated directly by counting the number of occurrences of each symbol in the whole text, and dividing them by the size in characters of the text. Similarly, the frequencies of occurrence of all possible lagged bigrams at 22 logarithmically spaced lags (powers of 1.5 rounded to integers, ranging from 1 to 7,482) were recorded, and the counts were divided by $N - n$, where N is the size of the corpus and n the desired lag, to obtain the probabilities $p^{(n)}(\sigma_i, \sigma_j)$ for all possible ordered pairs of symbols (σ_i, σ_j) in the alphabet (i.e., in Σ^2). From these raw estimates (the finite sample error will be dealt with in a further step), the lagged transinformations were computed by direct application of Eq. 14. These computations were performed separately for each individual corpus file (*i.e.*, session of the Parliament) in each of the four versions of the corpus, and for each file in each version I recorded the size of the file in characters (L), and the obtained estimates of the transinformation at the different lags $\hat{I}_n(L)$.

The estimates of I_n across the files (for each of the four versions of the corpus) were obtained using the linear extrapolation that is derived from Eq. 16. For each lag n, I performed a linear regression with the estimated values of $\hat{I}_n(L)$ across files as the dependent variable, and and the reciprocal of the corresponding file's length $(1/L)$ as the predictor. The intercept terms of these regressions were taken as the estimates of I_n corrected for the finite sample effects, with the standard errors of the intercept estimate providing the associated error bars.

It should be noted here that the alphabets observed across the files were slightly different. Larger files were more likely to contain further symbols than the alphanumeric characters and punctuation marks, for issues like currency markers, etc. This amounts to slight differences in the base of the logarithms of Eq. 14 across files, with the smaller files using slightly lower bases. In other words, the larger files were more accurate on average as they had a better approximation of the total alphabet size. To account for this, the contribution of each point to the regressions was weighted by the file size. The reliability of the method was assessed by inspecting the normality and stationarity of the regression residuals, as well as the quality of the fits.

Symbolic Fourier spectra. Each of the files was converted to 29 binary indicator sequences, each corresponding to one symbol in the simplified alphabet (see preprocessing). In order to have the same range of frequencies across all files, only the first 2^{15} characters in each file were considered.

For each indicator sequence, I computed the periodogram using the FFT algorithm. Prior to applying the FFT, the sequences were subjected to a linear detrending, and a split cosine bell taper was applied to the 10% of the data at the beginnings and ends of the sequences. The resulting periodograms were smoothed using three modified Daniell smoothers (*i.e.*, moving averages giving half weight to the end values) of widths 3, 5, and 7 (this whole procedure is implemented in the `spec.pgram` function of the *R* statistical package (R Development Core Team, 2005)). As indicated by Eq. 18, the overall symbolic spectrum for each file was computed as the sum of the powers in the resulting periodograms. This process was repeated for each of the 69 sessions (separately for the four variations of the corpora), and an averaged spectrum was computed across sessions (providing also an estimate of the standard error).

Lempel-Ziv complexity. For each file in each of the four variants of the English section of the *Europarl* corpus, I estimated Lempel-Ziv complexity using Eq. 20. The parsing was performed using an in-house implementation of the LZ76 algorithm (Lempel & Ziv, 1976). For each file, I computed values of L_n for values of n increasing in intervals of 10,000 characters (*i.e.,* I first considered the first 10,000 characters, then the first 20,000, 30,000, and so on up to the full length of the file).

The individual values of the asymptotic value L_∞ for each file were obtained using a two level estimation procedure based on Eq. 22. On a first stage, the values of the parameters $\hat{h}$, $\hat{a}$, $\hat{b}$ were estimated for the whole set of files (separately for each of the four versions of the corpus). The fit was peformed using a non-linear least squares procedure (implemented by the *R* function `nls`; R Development Core Team, 2005). The fits are illustrated in Fig. 1.

In the second stage, the exponent parameter $\hat{b}$ was assumed to remain approximately constant across the whole population, with the variation across files circumscribed to the other two parameters (that b is roughly constant across files was supported by separate individual fits to the data, and is also visible in Fig. 1). Assuming a constant value of $\hat{b}$, the individual estimates for each file F ($\hat{h}_F$ and $\hat{a}_F$) were obtained by a linear regression with the computed values of L for each file as the dependent variable, and the quantity $n^{-b} \ln n$ as the independent variable, so that the intercepts of these regressions correspond to the estimates of $\hat{h}_F$ and their slopes estimate $\hat{a}_F$.

The robustness and reliability of both levels of the regressions were assessed by inspecting the normality and stationarity of the model residuals, as well as the accuracy of the fits provided (which had explained variance levels $r^2 > 98\%$ in all cases). Notice that this multilevel approach to the fit can be interpreted as a non-linear random effects model enabling for random effect variation in the parameters a and h.

Results & discussion

Multifractal structure mimics linguistic description. Fig. 2a plots the scaling of the transinformation as a function of lag for the four versions of the English corpus. It can be clearly seen that the mutual information has a significant value – clearly above chance – extending up at least a few hundred items, well beyond the maximum of five words that was argued for by Huang et al. (1993) and Pothos and Juola (2007). I think that that supposed five word limit is nothing but a mirage created by computing the transinformations using words – instead of characters – as the basic unit of study, hence being much more prone to finite sample effects (which were also not corrected for in the studies of Huang et al. or Pothos and Juola). Furthermore, the clear power-law pattern that is signaled by the straight line components of the figure is consistent with the scaling previously observed by many (Li, 1989, 1990; Li & Kaneko, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Ebeling & Frömmel, 1998), and not with the exponential decay that was assumed *a priori* by Pothos and Juola (2007). As I will show below, this power law properties extend across languages and genres.

More importantly, as was suspected, the original file shows a piecewise linear pattern, as would be typical of a multifractal system. At least four different scaling regimes can be found in the plot. First, at lags between one and five characters, one finds a regime that is mostly common to all variants of the corpus, save for the version that is fully randomized on the level of letters. Hence, we can argue that this part of the curve reflects the regime found within words, that is, lexical information and morphological regularities (this is further supporting by examining the average word length in the corpus, which is between four and five characters per word).

Second, between five and roughly 20 characters of lag, one finds a marked straight line, which is present both in the original and in the sentence shuffled corpora, but not in the word or

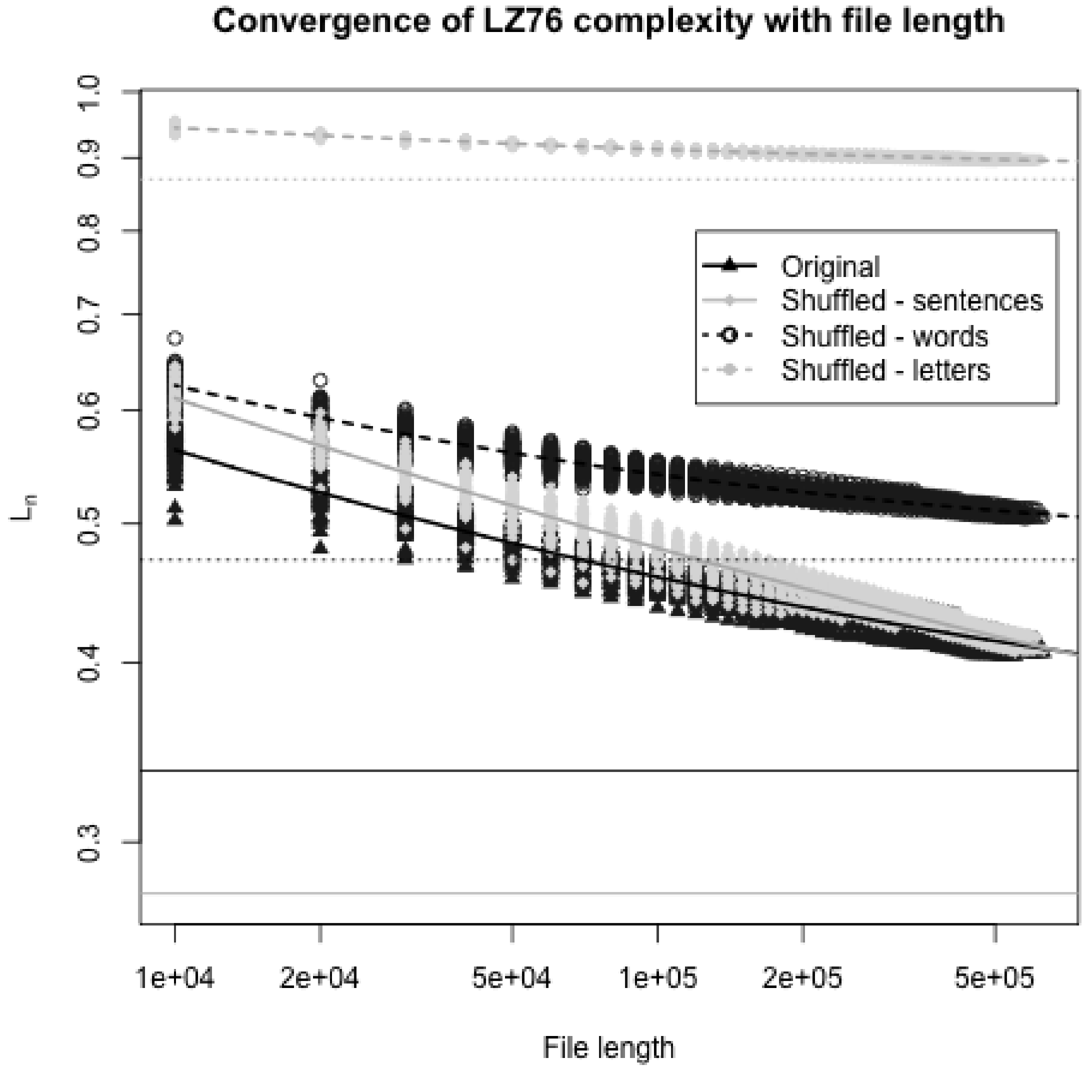

Figure 1. Convergence of the Lempel-Ziv complexity to its asymptotic value, and finite sample corrections. The dots represent the complexity estimates for each individual file at each individual length. The thicker lines represent the first stage of the fit. The thinner horizontal lines plot the mean asymptote for each group of files. Note the double logarithmic plane.

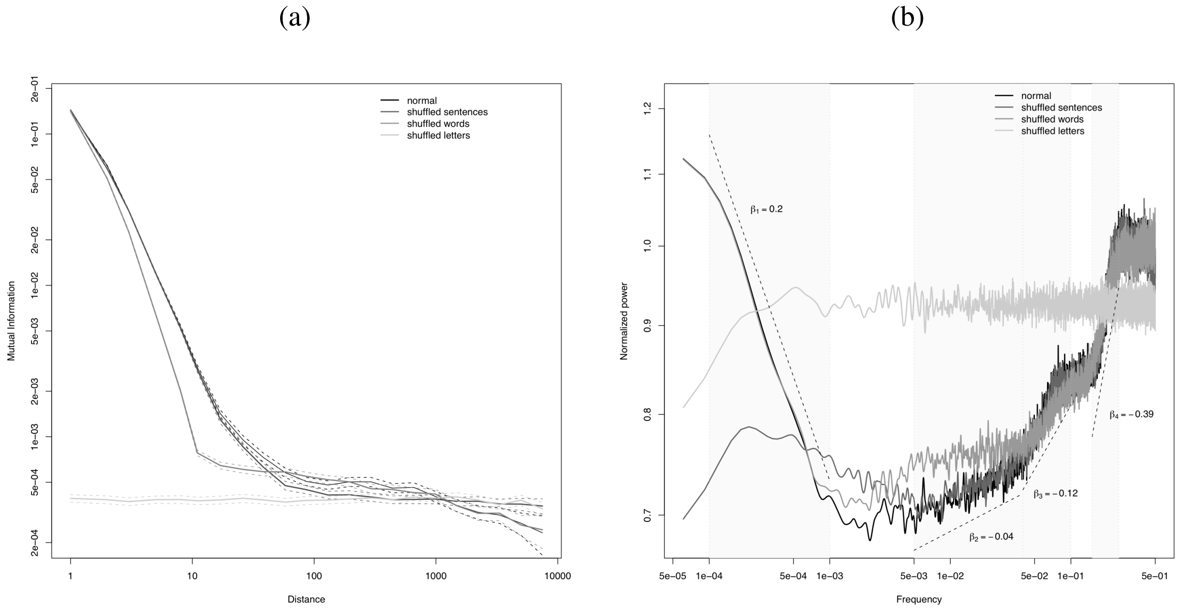

Figure 2. Multifractal scaling of texts. **(a)** Averaged transinformation spectra for the original texts, the texts randomized per sentence, randomized per word within sentence, and randomized per letter. The dashed lines plot one standard error above and below the averages. Note the logarithmic scale of both axes. **(b)** Averaged symbolic Fourier spectra for the original texts, the texts randomized per sentence, randomized per word within sentence, and randomized per letter. Note the logarithmic scale of both axes.

letter shuffled ones. This is indicative of this regime corresponding to syntactic relations. Similarly, a slightly different regime seems to be present between approximately 20 and 80 characters. This regime is also destroyed when shuffling by word or by letter, thus we can assume it is also syntactic in origin.

Finally, most clearly at lags between 100 and 1,000 characters, one finds a final regime, which is present in the original corpora and in the corpora shuffled by words, but not in the sentence or letter shuffled versions. What is common to the original and the word shuffled, but absent in the rest, is the pragmatic structure of the text, the topical ordering of sentences and paragraphs within the overall discourse.

The symbolic Fourier spectra plotted in Fig. 2b provide additional support for the interpretation above. As before, the spectrum exhibits a piecewise linear pattern, with four salient linear components (highlighted by the shadings in the graph). The regime labelled by the spectral slope β_1 corresponds roughly of the rightmost regime in the transinformation spectrum,[6] denoting the pragmatic structure in the discourse. Once more, this regime is common to the original and to the word shuffled versions of the corpora, but not to the sentence shuffled or the letter shuffled version.

To the left, one finds the regimes labelled with the regression slopes β_2 and β_3. These regimes are apparent both in the original files, and in the files shuffled by sentence, but not in those shuffled by word or by letter. Therefore, as was the case with he two corresponding regimes in the transinfor-

[6] The smoothing techniques used on the Fourier spectra result in a slight leftwards shift of the spectral components, resulting in a slight underestimation of the frequencies at which each regime begins and ends.

mation spectrum, these two scalings reflect the syntactic properties of language. I hypothesize that these two levels correspond respectively to the shallow syntactic structures at the phrase or clause level (β_3; between 5 and 20 characters length), and the deeper structures of sentences, building on the phrases and clauses (β_2; roughly between 20 and 80 characters).

Finally, the rightmost regime, labelled with the scaling β_4, is common to all versions of the corpora save for those randomized by letter. This is to say, as was the case with the leftmost regime in the transinformations, this reflects the scaling arising from lexical and/or morphological properties of the words.

The values of the spectral scaling exponents have direct interpretations. When $0 < |\beta| < 1$, the scalings correspond to Symbolic Fractional Gaussian Noises (SfGn; Li, 1990; Li & Kaneko, 1992). If $\beta > 0$, the noise is persistent, which is the equivalent of positive correlations between numerical data, indicating that an item tends to be similar to the preceding ones; if $\beta < 0$ the noise is anti-persistent, as with negative correlations this implies that the identity of consecutive symbols tends to be as different as possible.

I estimated the values of the scaling index β for each of the four regions identified in Fig. 2b. The estimate was obtained as the (minus) slope coefficient of a linear regression with log frequency predicting log power, including only the points in each of the regions identified (whose approximate limits are denoted by the dotted lines in the figure). The resulting indices are the four β values provided in the figure, and are illustrated by the associated dashed lines.

On the one hand, the components that are associated with lexical and syntactic structures (β_2, β_3, and β_4) are all negative, as is characteristic of anti-persistent SfGn's. This indicates that, at lexical and syntactic levels, the repetition of structures is discouraged. That is to say, it is relatively rare to observe the same letter repeated several times consecutively, or to observe same morpheme repeated within the same word, the same word multiple times within te same phrase, or the same type of phrase structure repeated within the same sentence.

On the other hand, β_1 – the scaling index that reflects pragmatic or discourse structure – is positive, as in a persistent SfGn. This indicates that on a long scale, there is a tendency for the repetition of similar structures. For instance, the topical ordering of discourse makes that the appearance of a word within a sentence is more likely if that word has appeared in recent sentences (but not within the same one), and the same can be expected of constructions, phrases, and syntactic structures across sentences.

In summary, I observed a decreasing value of the spectral index β with increasing scale. The larger scale (pragmatics) shows a tendency for persistence, with the finer grained scales becoming more anti-persistent. The level of anti-persistency increases as one decreases the scale, indicating that the constrains against repeating the same or very similar structures are stronger the smaller the context one is considering. As is evident from Fig. 2b, the gradation of the index is not progressive, but rather abrupt, hence the piecewise linear appearance of the graph, indicating the transition between the discrete levels of linguistic description. As I will discuss below, these findings are in good agreement with linguistic and psycholinguistic results.

A question that arises is to which extent the pattern observed is specific to the English language or, alternatively, to the particular discursive style of parliamentary sessions (especially so, taking into account that the European Parliament is famously peculiar in having speakers simultaneously using a multitude of languages within the same debate, all receiving simultaneous translation into many other languages; it is far from being a typical context of language use).

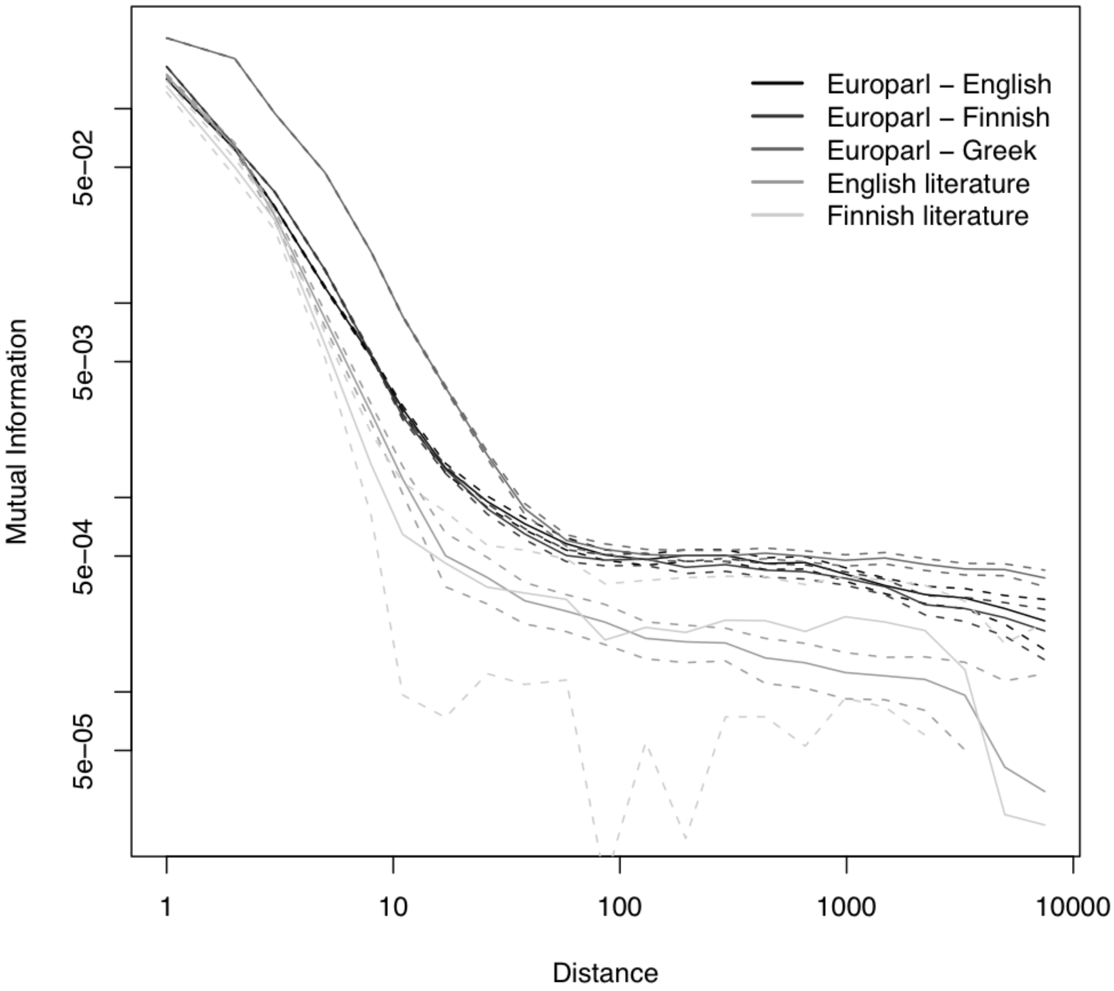

Figure 3. Transinformation spectra across styles and languages.

Across languages and genres. With regard to genre, a pattern practically identical to that of Fig. 2b has been found to describe the spectrum of literary texts (Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Pavlov et al., 2001; Gillet & Ausloos, 2008), suggesting that the overall structure observed may not be too dependent on the type of text. Fig. 3 compares the transinformation spectra of the English, Finnish, and Greek sections of the *Europarl* corpus, and the collections of English and Finnish literary texts. It is clear from the figure that the main clustering factor is not the actual language, but rather the genre, and in all cases the spectra contain the four phases identified above. With respect to linguistic differences, the Greek section of *Europarl* appears to stand out from its English and Finnish counterparts. Notice, however, that this difference arises mostly from Greek having significantly higher values of the transinformation at the very short lags. I think that this effect is mainly due to alphabetic differences. Unlike English and Finnish, which are transcribed using the Roman alphabet, Greek is transcribed using mostly the Greek alphabet. However, non-Greek personal names – which are evidently abundant during the Parliament sessions – remain in the Roman alphabet also in the Greek transcription. Therefore, the letters are clustered into two groups and, as the alphabets are never mixed within the same word, that the first letter in a word is written in one of the alphabets automatically sets to zero the probabilities of all letters in the other alphabet, increasing considerably the transinformations at short distances. Notice that beyond the word level, the slopes of the syntactic and pragmatic scaling regimes are basically identical in the three languages. Also at the lexical level, a fine but significant difference (see the error bars) distinguishes Finnish from English. The lower slope section at the leftmost part of the lines is slightly longer in Finnish than in English, possibly reflecting that – on average – Finnish words are longer than English one, mostly because of the richer agglutinating morphological system of Finnish. Once the pattern is established into the syntactic regimes, there are no significant differences between the three languages, neither on the *Europarl* corpus, nor between the two literary samples.

The differences are much more notable between genres, consistent with the findings of Genzel and Charniak (2003) and Keller (2004). However, one must also notice that the same four regimes are clearly visible in both sets, hence what is changing is the individual value of the scaling index for each of the levels of description, but the levels themselves remain the same. As one could expect, even between the genres, the lexical regime remains fairly unchanged, the differences becoming clear at the syntactic and pragmatic levels. This reflects differences in syntactic and discourse structure between literary texts and political debates, which is not all that much surprising, but it is interesting that it can already be detected from plain lagged transinformations between characters.

Adjustments between levels: from characters to words, and beyond. I now investigate whether adjustments in word length along the text can to some degree compensate for the decrease in the word-based conditional entropy that will result from the LRC property. The theoretical analysis showed that – strictly speaking – it is not possible to adjust the word lengths so that the word-based conditional entropy remains constant. However, Eq. 33 implies that a very slow increase in word length approaching an asymptote, with values of $\eta < 1$, gives rise to a non-monotonic evolution of the conditional entropy, with an initially decreasing phase, which is followed by a slowly increasing pattern from the point where the increase in word length dominates the decrease in character-based conditional entropy. Although not constant, the rate of increase of the word-based conditional entropy will, from this point onward, be extremely slow (and will be decreasing following a power law with exponent η), and could be considered as a constant for all practical matters, leading to an *approximately* valid instance of the rCER hypothesis at the level of words. As was discussed above, non-stationarity is part and parcel of the mutifractal structures, such as the ones I have described. It is thus reasonable to expect that the multifractal structure produces dynamical adjustments in word length.

The grey dots in Fig. 4 plot the average lengths of words starting at different positions in the text. To compute these, each of the English *Europarl* files was divided into blocks of 32 characters. In each block in each file, I computed the average length of the words that started within that block. For the early blocks, this provided 69 estimates of the mean word length within that block, and progressively fewer estimates for later blocks (*i.e.*, shorter files have less blocks). The dots plot the average word length of the blocks across all files, as a function of position in the text (*i.e.,* 32 × Block number). At least in the very early positions, there seems to be a marked increase in the mean word lengths, which could be in line with the proposed adjustments in word length. The parameters $l_{\min}$, Δ_l, and η from Eq. 32 were estimated by a least-squares non-linear regression, fitting the mean word lengths by position from all individual files. The regression revealed significant values of the three parameters ($\hat{l}_{\min} = 4.3296 \pm .1920$, $t[145793] = 22.555$, $p < .0001$, $\widehat{\Delta}_l = .5522 \pm .1698$, $t[145793] = 3.251$, $p < 0.0011$, and $\hat{\eta} = .2202 \pm .0709$, $t[145793] = 3.108$, $p < 0.0019$). The fitted regression is plotted by the solid line in the figure.

The estimated value of $\hat{\eta} = .2202 < 1$ confirms the hypothesized very slow increase in the word lengths, that could perhaps balance the per-word entropy rate to some degree. To ensure that estimated increase in word lengths was not solely due to the strong growth in the early positions, the estimation was repeated excluding all positions below 2,500 characters (dotted line in the figure), and the estimates remained significant even then. However, going further right than 2,500 characters made the increase in word length too small to be detectable by the regression (neither by the non-linear one above, nor by alternative linear and log-linear models allowing unbounded growth).

The above shows that there is indeed some degree of dynamical adjustment of the word lengths, reflecting a pressure to increase the efficiency in the use of the channel. Combining the estimates of $h = .07$, $g = 1/4$, and $\mu_1 = 1/2$ for English that are available in the literature (Ebeling & Frömmel, 1998; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Hilberg, 1990), with the estimated shape of the word length increase estimated above, one can use Eq. 33 to estimate the evolution of the per-word conditional entropy $h_w^{(n)}$. This estimate is plotted in Fig. 5. The figure shows the predicted non-monotonic pattern, with a relatively strong decrease in $h_w^{(n)}$ up to a global minimum, and a very slow asymptotic increase from there onwards. This could be interpreted as indicating that $h_w^{(n)}$ is to all practical effects constant from the point that the minimum is reached. Notice, however, that for these values of the parameters, the location of this minimum is in the order of $3 \cdot 10^5$ characters, that is, about the duration of a full long session of the European Parliament. This would mean that, for most human utterances (which are significantly shorter than that) the per-word conditional entropy would be decreasing along position, invalidating the rCER hypothesis.

These results show that speakers adjust their word lengths so as to improve their efficiency in communication, but this is still insufficient to achieve a constant entropy rate at the level of words. The theoretical predictions hold not only for the extension from characters to words, but more generally for the extension from characters to any larger linguistic unit. This is to say, the LRC property observed at the character level makes it impossible to keep a constant entropy rate at any linguistic level of description.

The multifractal properties of language described above suggest that speakers may be making

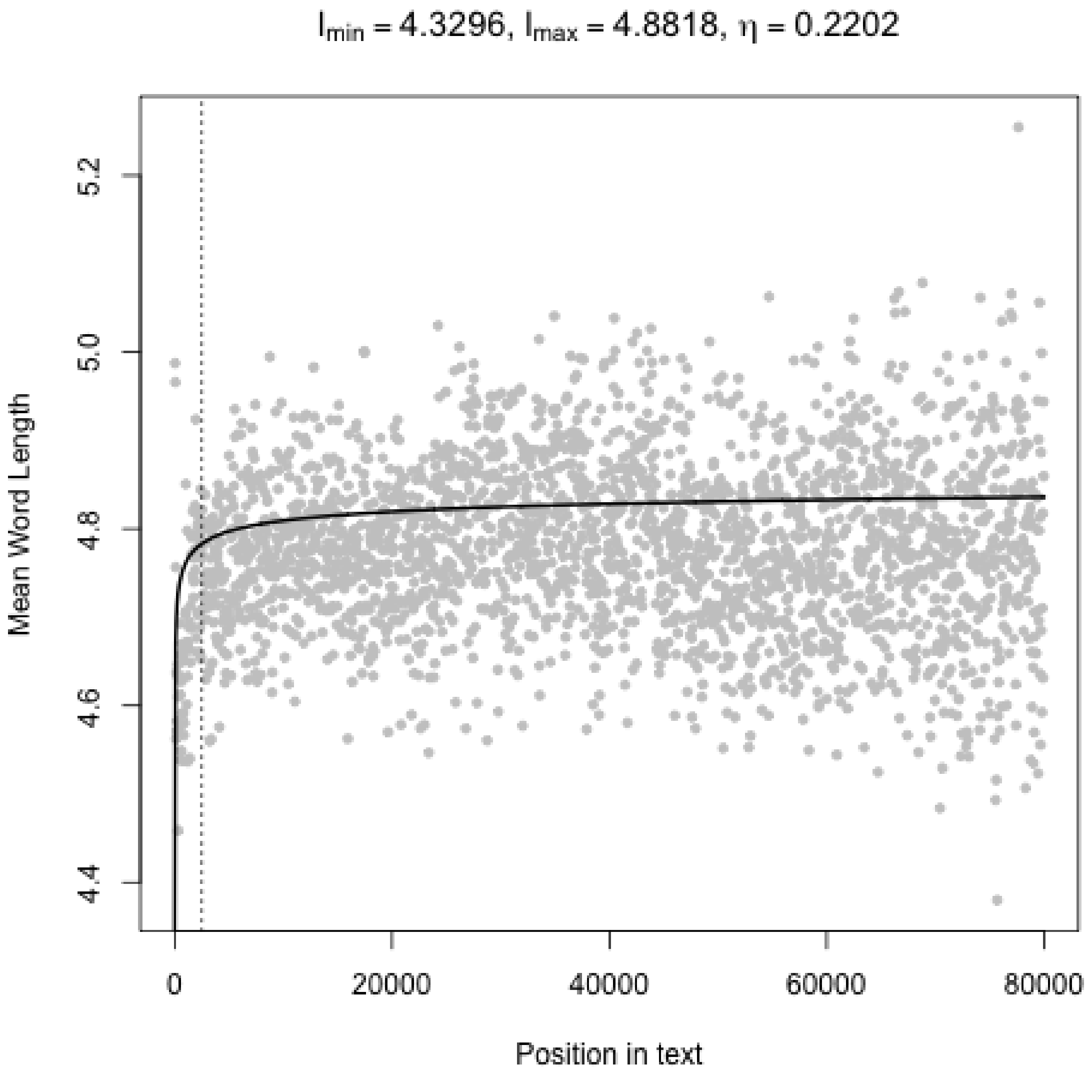

Figure 4. Increase of word length modeled as a function of the position of the word in the text. The grey dots plot the average word lengths at a position across files. The solid line plots a least-squares nonlinear regression fitting the parameters of Eq. 32 (performed on the individual values for each file).

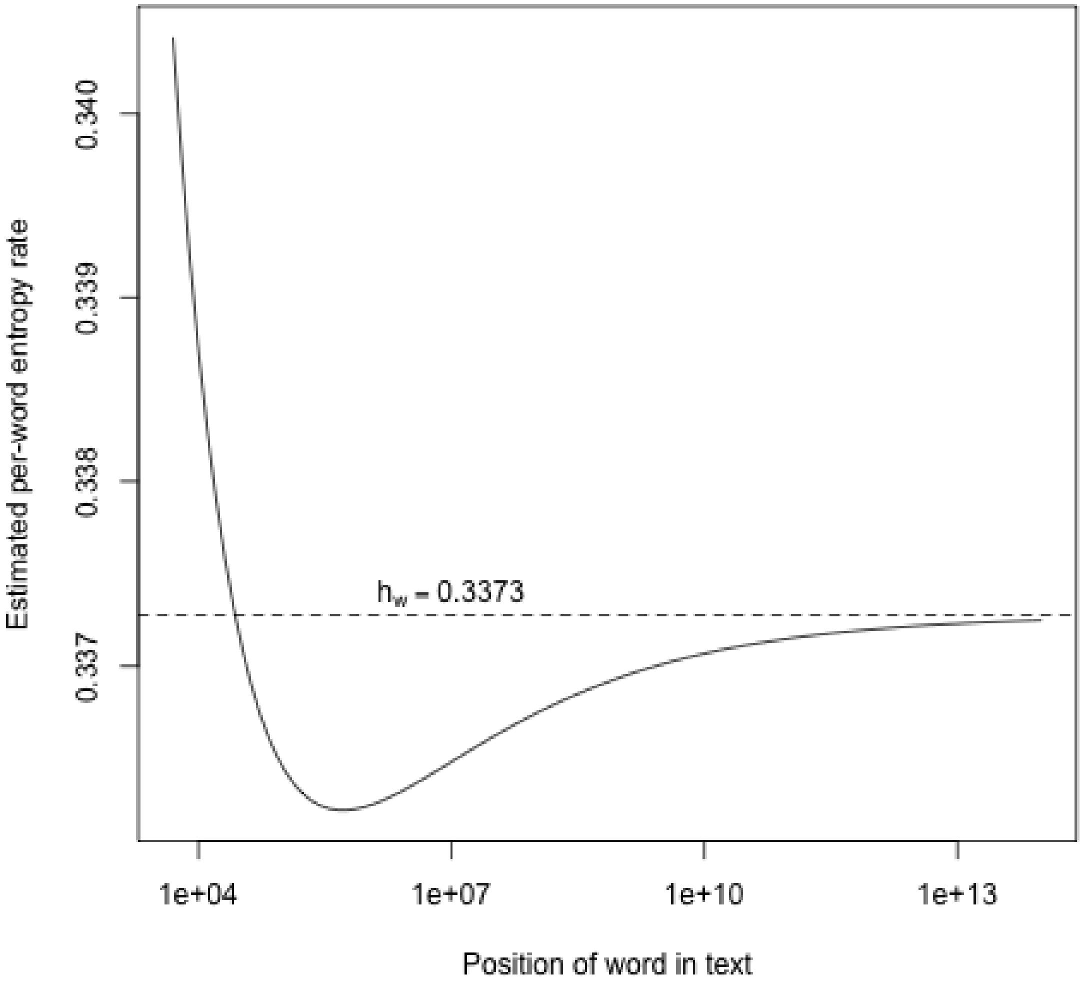

Figure 5. Estimated evolution of the conditional entropy per word through long English texts, assuming $h = .07$, $g = 1/4$, $\mu_1 = 1/2$ (Ebeling & Frömmel, 1998; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Hilberg, 1990), and using the growth in word length with position estimated from the corpora. The horizontal dashed line plots the estimated asymptote at infinite position, corresponding to the per-word entropy of the source (in $\log_\lambda$ units).

simultaneous adjustments at multiple levels in order to improve the efficiency of their channel use, and psycholinguistic experiments support this possibility (Aylett & Turk, 2004; Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007; McDonald & Shillcock, 2001; Piantadosi et al., 2009). For instance, in parallel to the increase in word length along textual positions, one also observes an increase in sentence length along a text (Keller, 2004; Qian & Jaeger, 2009). If one is simultaneously making adjustments at many levels, the adjustments at all levels must themselves be compensated for, to avoid adjustments in one level damaging adjustments in another too dramatically. For instance, if one is increasing the mean word length in characters, the increase in mean sentence length in words must at that point be attenuated to avoid resulting in overcorrections at the sentence level.

Notice that in Fig. 4, although the average non-stationary pattern follows the estimated regression line, there is still a very large amount of variability in the word lengths, that is not accounted for by the growth pattern. The dots in the figure actually correspond to averages across many files, and still they exhibit a significant spread above and below the main pattern. If this spread reflects the simultaneous presence of adjustments at many other nested levels, as would follow from the multifractal structure, the oscillations around the pattern must also themselves be of a fractal nature; their multi-scale coordination requires that the adjustments are correlated across all scales and lags. This implies that the residuals of the non-linear regression should exhibit a fractal pattern for each of the individual files. To test this hypothesis, I used the Bayesian Assessment of Scaling methodology (BAS; Moscoso del Prado Martín, 2010). The BAS compares two hypotheses on the scaling regime of a time series. For each individual file, I compared the hypothesis that the residuals exhibited at most short-term correlations, that is, the spectral index is $\beta \approx 0$, with the hypothesis of persistent LRC of the type found in fractional Gaussian noises (Mandelbrot & Van Ness, 1968), which have spectral scaling exponents $0 < \beta < 1$. Across the 69 files, I observed 1,396 decibels of evidence in favor of the LRC hypothesis, that is, on average about 20 decibels per file (ten decibels is already strong statistical evidence to prefer one hypothesis over another; cf., Moscoso del Prado Martín, 2010). The BAS also provides a maximum likelihood estimate of the spectral scaling index, which on average was $\hat{\beta} = .2786 \pm .006 > 0$, consistent with the persistent LRC hypothesis.

Summarizing, speakers appear to adjust the information processing rate simultaneously at multiple levels (word lengths, phrase lengths, sentence lengths, *etc*.), as is advocated by proponents of different versions of the CER hypothesis (Aylett & Turk, 2004; Fenk & Fenk-Oczlon, 1993; Fenk-Oczlon & Fenk, 1999, 2002; Frank & Jaeger, 2008; Genzel & Charniak, 2002, 2003; Jaeger, 2010; Keller, 2004; Levy & Jaeger, 2007; Manin, 2006; Piantadosi et al., 2009; Qian & Jaeger, 2009; Vega & Ward, 2009). These adjustments will improve the efficiency in the use of the channel from all perspectives, but the adjustment will never be sufficient for making the entropy rate constant at any of the given levels. I estimate that, despite the word length adjustments, the entropy rate is decreasing up to very advanced points in discourse, following the pattern illustrated by Fig. 5.

Decomposition of the linguistic levels by Lempel-Ziv complexity. The distribution of Lempel-Ziv complexity values for the four versions of the corpora is summarized in Fig. 6a. The average complexity of the original files was $\hat{h} = .3382 \pm .0027$. This value corresponds to $1.64 \pm .01$ bits per letter, and is thus in the range of the upper-bound estimates for the per-letter entropy of English available in the literature (Burton & Licklider, 1955; Brown, della Pietra, Mercer, della Pietra, & Lai, 1992; Cover & King, 1978; Moradi et al., 1998; Rosenfeld, 1996; Shannon, 1951). In contrast, the complexities of the version randomized by sentence was $\hat{h}_S = .2521 \pm .0020$, that of the version with the words shuffled within each sentence was $\hat{h}_W = .4728 \pm .0017$, and the complete

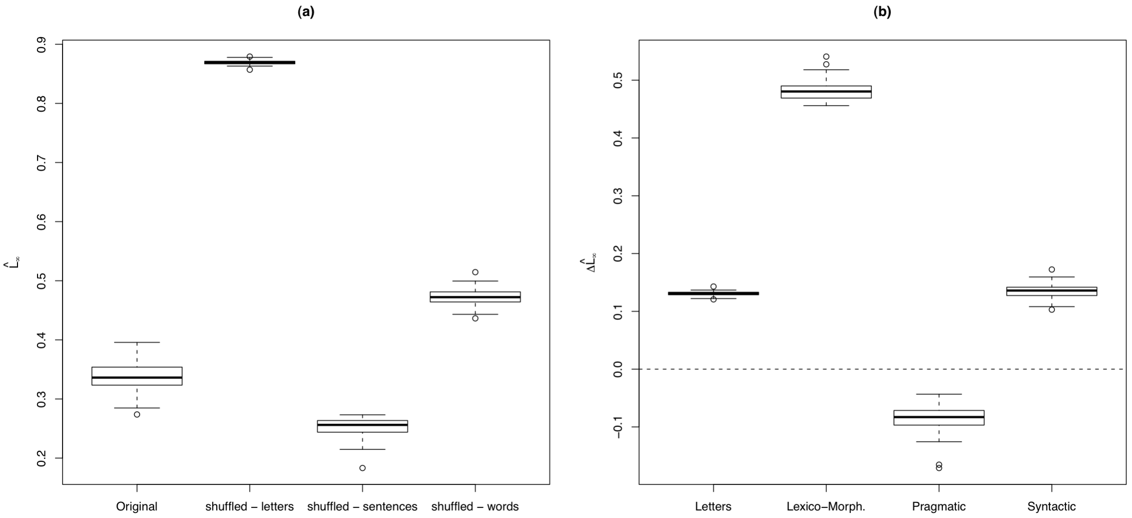

Figure 6. **(a)** Asymptotic Lempel-Ziv complexities for the four variants of the *Europarl* English corpus, across the 69 files studied. **(b)** Decomposition of the sources of information contributing to the entropy.

randomization by letters had a complexity of $\hat{h}_L = .8691 \pm .0004$, which is, as one would expect, very proximal to the maximum complexity of one, the difference being due to the non-uniform frequencies of the characters in the text. Indeed, estimating the entropy of the letter frequency distribution, one obtains an estimate of $\hat{h}_L = .8548$, confirming the accuracy of the asymptotic correction employed here. Following the approach described in Eq. 23, by design, each of the four versions of the corpus contains a different subset of information levels, leading to the system of linear equations,

$$\left.\begin{array}{lcl} 1-h & = & I_{\text{letters}} + I_{\text{lexical}} + I_{\text{syntax}} + I_{\text{pragmatics}} \\ 1-h_S & = & I_{\text{letters}} + I_{\text{lexical}} + I_{\text{syntax}} \\ 1-h_W & = & I_{\text{letters}} + I_{\text{lexical}} + I_{\text{pragmatics}} \\ 1-h_L & = & I_{\text{letters}} \end{array}\right\}, \tag{34}$$

whose solution is given by:

$$\left.\begin{array}{lcl} I_{\text{letters}} & = & 1-h_L \\ I_{\text{lexical}} & = & h+h_L-h_S-h_W \\ I_{\text{syntax}} & = & h_W-h \\ I_{\text{pragmatics}} & = & h_S-h \end{array}\right\}. \tag{35}$$

Applying these expressions to the Lempel-Ziv estimates of complexity, one obtains the estimates of Fig. 6b. The letter frequencies provide an estimated information of $\hat{I}_{\text{letters}} = .1308 \pm .0004$, the morphological and lexical information accounts for $\hat{I}_{\text{lexical}} = .4825 \pm .0021$, the syntactic information is of about $\hat{I}_{\text{syntax}} = .1346 \pm .0014$, and, most remarkably, the value of the information accounted for by the sentence ordering is negative $\hat{I}_{\text{pragmatics}} = -.0861 \pm .0027$.

That this value is negative is highly significant ($t[68] = -32.06$, $p < .0001$) and in no way an artifact of this particular dataset (I have replicated this same result across many different corpora, languages, and styles). It reflects the fact the randomizing the order of the sentences in a text *increases* the compressibility of the text. This is very counterintuitive, it implies that the compressibility of the text is better if some degree of additional randomness at long scales is added to it. This is at odds with the interpretation of the Lempel-Ziv complexity as an estimator of the entropy of the source. Increasing the randomness of the sequence must increase the entropy, but we are observing precisely the opposite. The explanation for this apparent paradox is that a precondition for the Lempel-Ziv complexity to converge to the entropy of the source is that the sequence is stationary. Hence, the negative change with increased randomness implies that this precondition is violated, so that the Lempel-Ziv complexity is not guaranteed to converge to the entropy of the source. This situation changes when one considers the results obtained on the randomized order sentences. In this case the randomization guarantees the stationarity of the source, and the convergence of the LZ76 algorithm is therefore guaranteed by the Ziv-Lempel theorems. In this case, the source entropy is estimated at $.2521$, that is, 1.22 bits per letter. This magnitude should correspond to an upper-bound to the entropy of English. The randomization of the sentence order can only have increased the entropy of the source. Therefore, although the non-stationarity does not enable us to estimate the real source entropy using LZ76, the randomization enables us to place a very precise upper bound. Indeed, that considered the best available estimate of the entropy of English, places it a precisely 1.23 bits per letter (Rosenfeld, 1996), which was obtained through a combination of multiple sources of linguistic information. Similarly, Cover and King (1978) report estimates ranging from 1.25 to 1.29 bits per letter depending on the text and method used. The results reported here indicate that those estimates are still upper bounds, so that the entropy rate of English is likely to fall within the bounds of $.6$ to 1.1 bits per letter estimated by Piotrovski (quoted by Levitin & Reingold, 1994).

As I discussed above, both the transinformation and Fourier spectra suggested also that the system is of a multifractal nature, and multifractal systems are non-stationary by definition (Ivanov et al., 1999). This is also supported by the observed increases in word length and sentence length along the text are also in contradiction with a stationary system. These results also suggest that the non-stationarity of texts is mostly governed by processes that work on relatively long scales, mainly the leftmost scaling regime in Fig. 2b. Multiple studies have suggested that long correlations are present between the letters of a text at many lags (Brown et al., 1992; Cover & King, 1978; Shannon, 1951), but some others have noted that, beyond 32 characters (Burton & Licklider, 1955; Moradi et al., 1998) or five words (Huang et al., 1993; Pothos & Juola, 2007), prediction in English is not significantly improved by considering further context. In fact, the results here indicate that the prediction accuracy could actually be damaged if significantly longer scales of context are considered in letter by letter prediction, if no adjustments are made for the non-stationary properties of texts.

General Discussion

My results confirm that linguistic sequences exhibit multifractal structure, in line with previous reports (Ebeling & Neiman, 1995; Ebeling, Neiman, & Pöschel, 1995; Gillet & Ausloos, 2008; Lardner et al., 1992; Pavlov et al., 2001). This is reflected in the piece-wise linear pattern of the transinformation and Fourier spectra in the double-logarithmic plane, by the significant non-stationarity of word lengths along a text, and by the non-convergence of the Lempel-Ziv complexity to the real value of the entropy rate. Furthermore, the choice randomizations employed, has revealed a direct correspondence between the linear components of the multiscaling spectrum and traditional levels of description used in Linguistics: Lexico-morphological, Syntactic (at shallow and deep levels), and Pragmatic. It is interesting that such levels already become apparent using theory-neutral tools such as the spectra. Notice here that, although the choice of randomizations cannot be taken as theory neutral (but it is still theoretically light), the regimes are clearly visible already in the untransformed spectrum. The randomizations have served to confirm the association with linguistic levels, but the different scales do not depend on it.

The importance of non-stationarity

A common working assumption of most statistical studies on human language is that it can be modeled as the result of a stationary ergodic process. This simplification is useful in the sense that it enables the use of most common probability estimation methods. However, the lagged transinformations, the Fourier spectra, the word length adjustments, and the Lempel-Ziv complexity analyses of the corpora have all pointed towards a fundamental non-stationarity of linguistic sequences. Importantly, the analyses have also shown that non-stationairity has important consequences on experimental results and could potentially lead to incorrect interpretations. Therefore, the results presented here indicate that much attention should be paid to either removing non-stationarities prior to statistical analyses of texts or, when this is not possible, explicitly considering how this could affect the results.

From antipersistence to persistence, consistent with psycholinguistic results

The scaling coefficients of the Fourier spectrum (the β_i) are themselves related forming what seems like a smooth function, as befits a multifractal process. At the longer – pragmatic – scales β_1 has a positive value, which decreases at increasing pace as one approaches the smallest lexical scale, going through different levels of syntactic structure. These values indicate a transition from persistent structures in the longer scales (elements are likely to be repeated or similar at short distance), to an increasingly anti-persistent structure at the shorter scales (elements are unlikely to be repeated or similar at short distance). In other words, the spectra reveal a pressure towards a smooth signal in the long scales (beyond the single sentence), and towards a rougher signal in the shorter scales (within a sentence or word).

This transition from anti-persistence to persistence is supported by a range of findings in the Psycholinguistics and Corpus Linguistics literatures. On the one hand, lexical priming experiments have shown that, when a word is presented shortly after an orthographically similar one, its recognition is slower and less accurate relative to controls. This phenomenon happens both when the prime is presented subliminally, as in *masked priming* (Bijeljac-Babic, Biardeau, & Grainger, 1997; Davis & Lupker, 2006; De Moor & Brysbaert, 2000; Drews & Zwitserlood, 1995; Grainger, Colé, & Segui, 1991; Grainger & Ferrand, 1994; Segui & Grainger, 1990), or consciously, as in *overt priming* (Colombo, 1986; Lupker & Colombo, 1994; Segui & Grainger, 1990). Also, in eye-movement studies, the presence of orthographic neighbors immediately preceding a word in a text is also found to harden the recognition of the word (Paterson, Liversedge, & Davis, 2009; Perea & Pollatsek, 1998; Pollatsek, Perea, & Binder, 1999). One could argue that, in priming experiments, repetition of the exact same item actually leads to facilitation, people become faster and more accurate. However, it has been shown that, in those situations, there is actually no overt separation of the items, the two instances are perceived as two parts of the same signal (Kanwisher, 1987), in what has been termed the *repetition blindness* effect. Further investigations of this effect reveal that, when a word is repeated within the same sentence, or a letter is repeated within the same word, people have more difficulties both to remember the sentence or word (and its items), or to produce it (Abrams, Dyer, & MacKay, 1996; Arnell & Jolicoeur, 1997; Epstein & Kanwisher, 1999; Hochhaus & Marohn, 1991; Humphreys, Besner, & Quinlan, 1988; Kanwisher, 1987; Kanwisher, Kim, & Wickens, 1996; Kanwisher & Potter, 1990; Luo & Caramazza, 1995; MacKay, Miller, & Schuster, 1994; Vokey & Allen, 2002).

On the other hand, corpus studies have repeatedly found that speakers tend to repeat the same syntactic structures in proximal sentences, even across speakers in dialog contexts (Calude & Miller, 2009; Estival, 1985; Kempen, 1977; Tannen, 1984, 1989; Schenkein, 1980; Weiner & Labov, 1983). Likewise, psycholinguists have widely documented the phenomenon of *syntactic priming*. Listeners and readers are faster and accurate at recognizing sentences and clauses whose syntactic structure replicates that of recently encountered clauses (Dooling, 1974; Frazier, Taft, Clifton, Roeper, & Ehrlich, 1984; Ledoux, Traxler, & Swaab, 2007; Mehler & Carey, 1967; Tooley, Traxler, & Swaab, 2009), and speakers have less difficulties in producing clauses whose structure is similar to other recently encountered or produced sentences (Bock, 1986, 1989; Bock & Loebell, 1990; Bock, Loebell, & Morey, 1992; Levelt & Kelter, 1982; Pickering & Braningan, 1999).

In sum, both the language production and the language comprehension systems appear to be optimized for producing and understanding a language is which repetition of similar items in a short scale is disfavored, but repetition of the same or similar items in larger scales is encouraged, which is precisely what is observed in the Fourier spectrum of texts. This suggests that the pressures at different scales are different. The shorter time scales are mostly influenced by low-level, perceptual constrains. The main goal here is to maximize the identifiability of each item in the sequence. Thus, it is advisable that consecutive elements should have as high a contrast between them as possible, partly to avoid issues like repetition blindness. However, when considering larger linguistic units, the pressure is not so much on recognition any longer, but more on higher cognition factors. Reusing the same words and syntactic structures minimizes the effort required for parsing and constructing sentences, as previous work can be reused. Hence what is sought is not to maximize contrast, but to reuse as much of the previous cognitive structures as is possible. Interestingly, this gradient from perception to cognition can be already detected on the simple statistics of the most basic units, and appears to be well organized in one single multifractal structure.

Back to the CER

An important conclusion from my analyses is that linguistic sequences are intrinsically non-stationary, that is, the statistical distribution of the symbols (at whichever level the symbols are considered) does not remain constant but rather changes throughout the sequence. This has crucial implications for the plausibility of the CER hypothesis. Had linguistic sequences been stationary, the choice of possible evolutions of the entropy rate is in fact rather limited, it can either be monotonically decreasing or remain constant. However, a constant entropy rate is only possible in the absence of LRC. As linguistic sequences do exhibit LRC, the CER would just not be possible. The non-stationarity provides space for some adjustments of the entropy rate so as to attenuate its monotonic decrease. However, as was illustrated by Fig. 5, even these adjustments are insufficient to make it a constant rate. Therefore, strictly speaking, the CER hypothesis remains invalid (as well as the relaxed version introduced here as the rCER). In this respect, in this study I have introduced a parametric model capable of predicting the actual shape of the evolution of the entropy rate in the linguistic system. The extension of the model to other non-stationary trends, such as the observed increase in average sentence length along a corpus (Keller, 2004), is just a trivial application of exactly the same formulation that I introduced here.

Even if the CER as such is not attainable, there is ample psychological evidence that humans try to optimize their rate of information transmission in linguistic sequences (Aylett & Turk, 2004; Frank & Jaeger, 2008; Jaeger, 2010; Levy & Jaeger, 2007; McDonald & Shillcock, 2001; Piantadosi et al., 2009). In this respect, the analysis of the increase in word lengths revealed that, beyond the monotonic increase pattern, the residuals from this trend exhibit a complex pattern of interlocked adjustments at many levels, and in fact most of the variability in the word lengths is accounted for, not by the overall trend, but by the fractal pattern. This fractal pattern indeed seems consistent with simultaneous online optimization processes at multiple linguistic scales (Fenk & Fenk-Oczlon, 1993; Fenk-Oczlon & Fenk, 1999, 2002, 2002, 2005, 2007).

Optimal communication rather than optimal transmission

This brings us back to the question of optimality. The original proposal of the CER hypothesis (Genzel & Charniak, 2002) was based on an assumption of rationality on the use of the channel. This is, it assumes that the main goal in language production is to transmit a signal in the most effective possible way. Under such assumption, one can infer from Shannon's coding theorem (Shannon, 1948) that the optimal would be to transmit information approximating in as much as possible the channel capacity C. Hence, one can expect language to convey information at some constant rate $r \leq C$. This has however some rather counterintuitive implications. The constant entropy rate implies a system with extensive entropy, that is, the entropy of a full signal would grow linearly along the message. As discussed above, such are chaotic, or at most Markovian systems. In other words, this assumption would predict uncorrelated signals, the human mind would be working as a near-optimal data compressor. Language would then resemble an extremely compact signal with no correlations and little redundancy. But the study of language reveals that one of its most salient characteristics is precisely the pervading presence of redundancy at all levels.

Although important, the maximization of the information transmission rate is not the only constrain that the language producer must optimize. The first important thing to notice is that the goal of producing language is not to convey a signal through a channel as efficiently as possible. Rather, the system must be optimized to maximize communication, and to do so in a context with severe limitations. It has been shown that major statistical properties of language, such as Zipf's Law, arise from optimization processes that consider multiple constrains. Among these is indeed the efficiency in channel use, but also the efficiency in the correspondence between linguistic signals and their 'meanings' (Ferrer i Cancho, 2005a, 2005b, 2005c, 2005d; Ferrer i Cancho & Díaz-Guilera, 2007; Ferrer i Cancho & Solé, 2003), avoiding issues such as excessive use of synonymy (Ferrer i Cancho, 2005d; Ferrer i Cancho & Díaz-Guilera, 2007; Manin, 2008). In other words, what should be optimized is not the transmission of a linguistic signal, but rather, the transmission of the message it conveys. The message conveyed and the linguistic signals are not the same thing. In the words of Shannon (1948, p. 327):

> Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages.

What is not relevant to the engineer, whose goal is to transmit the very signal he/she was given, is perhaps the most crucial aspect for the human producing language, to convey the actual meaning of the message. Therefore, linguistic communication must be optimized for the latter rather than the former, although optimization of communication does involve some optimization of the transmission of the linguistic sequence.

Furthermore, neither speakers nor listeners are perfect systems, memory resources are limited and exhibit a temporal decay, which implies that at some point, both the speaker and the listener may loose track of the long correlations present in language, and thus not be able to use them to full effect. Rather than rationality, a more suggestive approach is what some have termed bounded rationality (Kahneman, 2003; Simon, 1991), that is, humans are rational, but within the bounds of their intrinsic limitations, which sometimes make it impossible to achieve optimal solutions. Nevertheless, precisely the nonextensive entropy of linguistic sequences, that is, the decay of the entropy rate with an exponent of approximately $1/2$ (Dębowski, 2006; Ebeling & Nicolis, 1991, 1992; Ebeling & Pöschel, 1994; Ebeling, Pöschel, & Albrecht, 1995; Hilberg, 1990) has been considered as evidence for language belonging to a class of systems referred to as *Highly Optimized Tolerance* (Saakian, 2005); these are basically the most efficient means of information transmission under complex restrictions.

References

Abrams, J. R., Dyer, J. R., & MacKay, D. G. (1996). Repetition blindness interacts with syntactic grouping in rapidly presented sentences. *Psychological Science*, *7*, 100–104.

Afreixo, V., Ferreira, P. J. S. G., & Santos, D. (2004). Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, *14*, 523–530.

Altmann, G. (1980). Prolegomena to Menzerath's Law. *Glottometrika*, *2*, 1–10.

Amigó, J. M., Szczpański, J., Wajnryb, E., & Sánchez-Vives, M. (2004). Estimating the entropy rate of spike trains via Lempel-Ziv complexity. *Neural Computation*, *16*, 717—736.

Amit, M., Shmerler, Y., Eisenberg, E., Abraham, M., & Shnerb, N. (1994). Language and codification dependence of long-range correlations in texts. *Fractals*, *2*, 7–13.

Arnell, K. M., & Jolicoeur, P. (1997). Repetition blindness for pseudoobject pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 999–1013.

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language & Speech*, *47*, 31–56.

Bijeljac-Babic, R., Biardeau, A., & Grainger, J. (1997). Masked orthographic priming in bilingual word recognition. *Memory & Cognition*, *25*, 447–457.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Bock, J. K. (1989). Closed class immanence in sentence production. *Cognition*, *31*, 163–186.

Bock, J. K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.

Bock, J. K., Loebell, H., & Morey, R. (1992). From conceptual roles to structural relations: Bridging the syntactic cleft. *Psychological Review*, *99*, 150–171.

Brown, P. F., della Pietra, V. J., Mercer, R. L., della Pietra, S. A., & Lai, J. C. (1992). An estimate of an upper bound for the entropy of English. *Computational Linguistics*, *18*, 31–40.

Burton, N., & Licklider, J. (1955). Long-range constraints in the statistical structure of printed English. *American Journal of Psychology*, *68*, 650–653.

Calude, A., & Miller, S. (2009). Are clefts contagious in conversation? *English Language and Linguistics*, *13*, 127–132.

Colombo, L. (1986). Activation and inhibition with orthographically similar words. *Journal of Experimental Psychology: Human Perception and Performance*, *12*, 226–234.

Cover, T., & King, R. (1978). A convergent gambling estimate of the entropy of English. *IEEE Transactions in Information Theory*, *24*, 413–421.

Crutchfield, J. P., & Feldman, D. P. (2003). Regularities unseen, randomness observed: The entropy convergence hierarchy. *Chaos*, *15*, 25—54.

Davis, C. J., & Lupker, S. J. (2006). Masked inhibitory priming in English: Evidence for lexical inhibition. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 668–687.

De Moor, W., & Brysbaert, M. (2000). Neighborhood-frequency effects when primes and targets are of different lengths. *Psychological Research*, *63*, 159–162.

Dębowski, Ł. (2006). On Hilberg's Law and its links with Guiraud's Law. *Journal of Quantitative Linguistics*, *13*, 81—109.

Dooling, D. J. (1974). Rhythm and syntax in sentence perception. *Journal of Verbal Learning and Verbal Behavior*, *13*, 255–264.

Drews, E., & Zwitserlood, P. (1995). Morphological and orthographic similarity in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *21*, 1098–1116.

Ebeling, W. (1997). Prediction and entropy of nonlinear dynamical systems and symbolic sequences with LRO. *Physica D: Nonlinear Phenomena*, *109*, 42–52.

Ebeling, W., & Frömmel, C. (1998). Entropy and predictability of information carriers. *Biosystems*, *46*, 47–55.

Ebeling, W., & Neiman, A. (1995). Long-range correlations between letters and sentences in texts. *Physica A: Statistical and Theoretical Physics*, *215*, 233–241.

Ebeling, W., Neiman, A., & Pöschel, T. (1995). Dynamic entropies, long–range correlations and fluctuations in complex linear structures. In M. Suzuki (Ed.), *Coherent approach to fluctuations (Proceedings of the Hayashibara forum 1995).* Singapore: World Scientific. Available from `http://arxiv.org/abs/adap-org/9507007`

Ebeling, W., & Nicolis, G. (1991). Entropy of symbolic sequences: the role of correlations. *EPL (Europhysics Letters)*, *14*, 191–196.

Ebeling, W., & Nicolis, G. (1992). Word frequency and entropy of symbolic sequences: a dynamical perspective. *Chaos, Solitons & Fractals*, *2*, 635–650.

Ebeling, W., & Pöschel, T. (1994). Entropy and long-range correlations in literary English. *EPL (Europhysics Letters)*, *26*, 241–246.

Ebeling, W., Pöschel, T., & Albrecht, K. (1995). Entropy, transinformation and word distribution of information-carrying sequences. *International Journal of Bifurcation and Chaos*, *5*, 51–61. Available from `http://arxiv.org/abs/cond-mat/0204045`

Epstein, R., & Kanwisher, N. (1999). Repetition blindness for locations: Evidence for automatic spatial coding in an RSVP task. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1855–1866.

Estival, D. (1985). Syntactic priming of the passive in English. *Text*, *5*, 7–21.

Fenk, A., & Fenk-Oczlon, G. (1993). Menzerath's Law and the constant flow of linguistic information. In R. Köhler & B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 11–31). Dordrecht, The Netherlands: Kluwer Academic.

Fenk-Oczlon, G., & Fenk, A. (1999). Cognition, quantitative linguistics, and systemic typology. *Linguistic Typology*, *3*, 151–177.

Fenk-Oczlon, G., & Fenk, A. (2002). The clausal structure of linguistic and pre-linguistic behavior. In T. Givon & B. F. Malle (Eds.), *The evolution of language out of pre-language* (pp. 215–229). Amsterdam, The Netherlands: John Benjamins.

Fenk-Oczlon, G., & Fenk, A. (2005). Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In G. Fenk-Oczlon & C. Winkler (Eds.), *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler* (pp. 75–86). Tübingen, Germany: Narr.

Fenk-Oczlon, G., & Fenk, A. (2007). Complexity trade-offs between the subsystems of language. In M. Miestamo, K. Sinnemäki, & F. Karlsson (Eds.), *Language complexity: Typology, contact, change* (pp. 43–65). Amsterdam, The Netherlands: John Benjamins.

Ferrer i Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. *Physica A*, *345*, 275–284.

Ferrer i Cancho, R. (2005b). Hidden communication aspects in the exponent of Zipf's law. *Glottometrics*, *11*, 96–117.

Ferrer i Cancho, R. (2005c). The variation of Zipf's law in human language. *The European Physical Journal B - Condensed Matter and Complex Systems*, *44*, 249–257.

Ferrer i Cancho, R. (2005d). Zipf's law from a communicative phase transition. *The European Physical Journal B - Condensed Matter and Complex Systems*, *47*, 449–457.

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, P06009–P06009.

Ferrer i Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 788 –791.

Frank, A. F., & Jaeger, T. F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 933–938). Austin, TX: Cognitive Science Society.

Frazier, L., Taft, L., Clifton, C., Roeper, T., & Ehrlich, K. (1984). Parallel structure: A source of facilitation in sentence comprehension. *Memory & Cognition*, *12*, 421–430.

Genzel, D., & Charniak, E. (2002). Entropy rate constancy in text. In P. Isabelle (Ed.), *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 199–206). Morristown, NJ: Association for Computational Linguistics.

Genzel, D., & Charniak, E. (2003). Variation of entropy and parse trees of sentences as a function of the sentence number. In M. Collins & M. Steedman (Eds.), *Proceedings of the conference on empirical methods in natural language processing* (pp. 65–72). Sapporo, Japan: Association for Computational Linguistics.

Gillet, J., & Ausloos, M. (2008). *A comparison of natural (English) and artificial (Esperanto) languages. a multifractal method based analysis.* (arXiv 0801.2510 [cs.CL])

Grainger, J., Colé, P., & Segui, J. (1991). Masked morphological priming in visual word recognition. *Journal of Memory and Language*, *30*, 370–384.

Grainger, J., & Ferrand, L. (1994). Phonology and orthography in visual word recognition: Effects of masked homophone primes. *Journal of Memory and Language*, *33*, 218–233.

Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Physics Letters A*, *128*, 369–373.

Herzel, H. (1988). Complexity of symbol sequences. *Systems Analysis, Modelling, Simulation*, *5*, 435–444.

Herzel, H., Schmitt, A., & Ebeling, W. (1994). Finite sample effects in sequence analysis. *Chaos, Solitons & Fractals*, *4*, 97–113.

Hilberg, W. (1990). Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannonschen Experimente? ("The well-known limit of redundancy free information in texts – a misinterpretation of Shannon's experiments?"). *Frequenz*, *44*, 243–248. (in German)

Hochhaus, L., & Marohn, K. M. (1991). Repetition blindness for pseudoobject pictures. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 422–432.

Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., & Rosenfeld, R. (1993). The SPHINX-II speech recognition system: An overview. *Computers, Speech and Language*, *2*, 137–148.

Humphreys, G. W., Besner, D., & Quinlan, P. T. (1988). Event perception and the word repetition effect. *Journal of Experimental Psychology: General*, *117*, 51–67.

Ivanov, C., Rosenblum, M. G., Amaral, L. A. N., Struzik, Z. R., Havlin, S., Goldberger, A. L., et al. (1999). Multifractality in human heartbeat dynamics. *Nature*, *399*, 461–465.

Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*, 23–62.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative 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.

Juola, P., Bailey, T. M., & Pothos, E. M. (1998). Theory-neutral system regularity measurements. In *Proceedings of the 20th annual conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *The American Economic Review*, *93*, 1449–1475.

Kanwisher, N. (1987). Repetition blindness: Type recognition without token individuation. *Cognition*, *27*, 117–143.

Kanwisher, N., Kim, J.-W., & Wickens, T. D. (1996). Signal detection analyses of repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1249–1260.

Kanwisher, N., & Potter, M. (1990). Repetition blindness: Levels of processing. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 30–47.

Keller, F. (2004). The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In D. Lin & D. Wu (Eds.), *Proceedings of the conference on empirical methods in natural language processing* (pp. 317–324). Barcelona, Spain: Association for Computational Linguistics.

Kello, C. T., Anderson, G. G., Holden, J. G., & Van Orden, G. C. (2008). The pervasiveness of $1/f$ scaling in speech reflects the metastable basis of cognition. *Cognitive Science*, *32*, 1217–1231.

Kempen, G. (1977). Conceptualizing and formulating in sentence production. In S. Rosenbaum (Ed.), *Sentence production: Developments in research and theory* (pp. 259–274). New York, NY: Erlbaum.

Khinchin, A. I. (1957). *Mathematical foundations of informarion theory*. New York, NY: Dover.

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/`

Kokol, P., & Podgorelec, V. (2000). Complexity and human writings. *Complexity International*, *7*. Available from `http://www.complexity.org.au/ci/vol07/kokol01/`

Lardner, C., Desaulniers-Soucy, N., Lovejoy, S., Schertzer, D., Braun, C., & Lavallée, D. (1992). Universal multifractal characterization and simulation of speech. *International Journal of Bifurcation and Chaos*, *3*, 715–719.

Ledoux, K., Traxler, M. J., & Swaab, T. Y. (2007). Syntactic priming in comprehension. *Psychological Science*, *18*, 135–143.

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

Lesne, A., Blanc, J.-L., & Pezard, L. (2009). Entropy estimation of very short symbolic sequences. *Physical Review E*, *79*, 046208.

Levelt, W., & Kelter, S. (1982). Surface form and memory in question answering. *Cognitive Psychology*, *14*, 78–106.

Levitin, L. B., & Reingold, Z. (1994). Entropy of natural languages: Theory and experiment. *Chaos, Solitons & Fractals*, *4*, 709–743.

Levy, R., & Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 849–856). Cambridge, MA: MIT Press.

Li, W. (1989). *Mutual information functions of natural language texts* (Working Paper Nos. SFI–89–008). Santa Fe, NM: Santa Fe Institute.

Li, W. (1990). Mutual information functions versus correlation functions. *Journal of Statistical Physics*, *60*, 823–837.

Li, W., & Kaneko, K. (1992). Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence. *EPL (Europhysics Letters)*, *17*, 655–660.

Luo, C., & Caramazza, A. (1995). Repetition blindness under minimum memory load: Effects of spatial and temporal proximity and the encoding effectiveness of the Þrst item. *Perception and Psychophysics*, *57*, 1053–1064.

Lupker, S. J., & Colombo, L. (1994). Inhibitory effects in form priming – evaluating a phonological competition explanation. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 437–451.

MacKay, D. G., Miller, M. D., & Schuster, S. P. (1994). Repetition blindness and aging: Evidence for a binding deficit involving a single, theoretically specified connection. *Psychology and Aging*, *9*, 251–258.

Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, *10*, 422–437.

Manin, D. Y. (2006). Experiments on predictability of word in context and information rate in natural language. *Informacionnye Processy*, *6*, 229—236.

Manin, D. Y. (2008). Zipf's law and avoidance of excessive synonymy. *Cognitive Science*, *32*, 1075–1098.

Maragos, P., & Potamianos, A. (1999). Fractal dimensions of speech sounds: Computation and application to automatic speech recognition. *Journal of the Acoustical Society of America*, *105*, 1925–1932.

McDonald, S. A., & Shillcock, R. C. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language & Speech*, *44*, 295–322.

Mehler, J., & Carey, P. W. (1967). Role of surface and base structure in the perception of sentences. *Journal of Verbal Learning and Verbal Behavior*, *6*, 335–338.

Moradi, H., Roberts, J. A., & Grzymala-Busse, J. W. (1998). Entropy of English text: Experiments with humans and a machine learning system based on rough sets. *Information Science*, *104*, 31–47.

Moscoso del Prado Martín, F. (2010). *Hypothesis testing on the fractal structure of behavioral sequences: the Bayesian Assessment of Scaling methodology.* Centre National de la Recherche Scientifique, Lyon, France. (*(manuscript submitted for publication)*)

Paterson, K. B., Liversedge, S. P., & Davis, C. J. (2009). Inhibitory neighbor priming effects in eye movements during reading. *Psychonomic Bulletin & Review*, *16*, 43–50.

Pavlov, A. N., Ebeling, W., Molgedey, L., Ziganshin, A. R., & Anishchenko, V. S. (2001). Scaling features of texts, images and time series. *Physica A: Statistical Mechanics and its Applications*, *300*(1-2), 310–324.

Perea, M., & Pollatsek, A. (1998). The effects of neighborhood frequency in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 767–779.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2009). The communicative lexicon hypothesis. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 2582–2587). Austin, TX: Cognitive Science Society.

Pickering, M., & Braningan, H. P. (1999). Syntactic priming in language production. *Trends in Cognitive Sciences*, *3*(4), 136–141.

Pickover, C. A., & Khorasani, A. (1986). Fractal characterization of speech waveform graphs. *Computers & Graphics*, *10*, 51–61.

Pollatsek, A., Perea, M., & Binder, K. (1999). The effects of neighborhood size in reading and lexical decision. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1142–1158.

Pothos, E. M., & Juola, P. (2007). Characterizing linguistic structure with mutual information. *British Journal of Psychology*, *98*, 291–304.

Qian, T., & Jaeger, T. F. (2009). Evidence for efÞcient language production in chinese. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 851–856). Austin, TX: Cognitive Science Society.

R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modelling. *Computers, Speech and Language*, *10*, 187–228.

Saakian, D. B. (2005). Error threshold in optimal coding, numerical criteria, and classes of universalities for complexity. *Physical Review E*, *71*, 016126.

Sabanal, S., & Nakagawa, M. (1996). The fractal properties of vocal sounds and their application in the speech recognition model. *Chaos, Solitons & Fractals*, *7*(11), 1825–1843.

Schenkein, J. (1980). A taxonomy for repeating action sequences in natural conversation. In B. Butterworth (Ed.), *Language production* (Vol. 1, pp. 21–47). London, England: Academic Press.

Schenkel, A., Zhang, J., & Zhang, Y. (1993). Long range correlations in human writings. *Fractals*, *1*, 47-55.

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

Segui, J., & Grainger, J. (1990). Priming word recognition with orthographic neighbors: Effects of relative prime-target frequency. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 65–76.

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

Shannon, C. E. (1951). Prediction and entropy of written English. *Bell System Technical Journal*, *30*, 50–64.

Silverman, B. D., & Linsker, R. (1986). A measure of DNA periodicity. *Journal of Theoretical Biology*, *118*, 295–300.

Simon, H. (1991). Bounded rationality and organizational learning. *Organization Science*, *2*, 125-134.

Stanley, H. E., & Meakin, P. (1988). Multifractal phenomena in physics and chemistry [review]. *Nature*, *335*, 405–409.

Stoffer, D. S., Tyler, D. E., & McDougall, A. J. (1993). Spectral analysis for categorical time-series: Scaling and the spectral envelope. *Biometrika*, *80*, 611–622.

Tannen, D. (1984). *Conversational style: Analyzing talk among friends*. Norwood, NJ: Ablex.

Tannen, D. (1989). *Talking voices: Repetition, dialogue and imagery in conversational discourse*. Cambridge, England: Cambridge University Press.

Tooley, K. M., Traxler, M. J., & Swaab, T. Y. (2009). Electrophysiological and behavioral evidence of syntactic priming in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 19–45.

Vega, A., & Ward, N. G. (2009). *Looking for entropy rate constancy in spoken dialog* (Tech. Rep. No. UTEP-CS-09-19). El Paso, TX: University of Texas at El Paso.

Vokey, J. R., & Allen, S. W. (2002). Repetition deficits, list context, and word-class interactions in the RSVP of words in sentences. *Canadian Journal of Experimental Psychology*, *56*, 98–111.

Voss, R. F. (1992). Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Physical Review Letters*, *68*, 3805–3808.

Voss, R. F., & Clarke, J. (1975). $1/f$ noise in music and speech. *Nature*, *258*, 317.

Wang, W., & Johnson, D. (2002). Computing linear transforms of symbolic signals. *IEEE Transactions on Signal Processing*, *50*, 628–634.

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*, 29–58.

Ziv, J., & Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, *23*, 337–343.

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