Patterns in syntactic dependency networks

Article *in* Physical Review E · June 2004

DOI: 10.1103/PhysRevE.69.051915 · Source: PubMed

3 authors, including:

Ramon Ferrer-i-Cancho
Universitat Politècnica de Catalunya
139 PUBLICATIONS **5,772** CITATIONS
SEE PROFILE

Ricard Sole
University Pompeu Fabra
412 PUBLICATIONS **19,277** CITATIONS
SEE PROFILE

Patterns in syntactic dependency networks

Ramon Ferrer i Cancho,[1,*] Ricard V. Solé,[1,2] and Reinhard Köhler[3]
[1]*ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain*
[2]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
[3]*Universität Trier, FB2/LDV, D-54268 Trier, Germany*
(Received 19 June 2003; revised manuscript received 19 September 2003; published 26 May 2004)

Many languages are spoken on Earth. Despite their diversity, many robust language universals are known to exist. All languages share syntax, i.e., the ability of combining words for forming sentences. The origin of such traits is an issue of open debate. By using recent developments from the statistical physics of complex networks, we show that different syntactic dependency networks (from Czech, German, and Romanian) share many nontrivial statistical patterns such as the small world phenomenon, scaling in the distribution of degrees, and disassortative mixing. Such previously unreported features of syntax organization are not a trivial consequence of the structure of sentences, but an emergent trait at the global scale.

DOI: 10.1103/PhysRevE.69.051915 PACS number(s): 87.10.+e, 89.75.−k, 89.20.−a

I. INTRODUCTION

There is no agreement on the number of languages spoken on Earth, but estimates are in the range from 3000 to 10 000 [1]. World languages exhibit a vast array of structural similarities and differences. Syntax is a trait common to all human languages and the subject of the present paper. More precisely, we aim at finding new statistical patterns in syntax. General statistical regularities that human language obeys at different scales are known [2–5]. Probably, the most striking regularity is Zipf's law for word frequencies [2]. Unfortunately, such a law seems to have nothing to do with syntax and symbolic reference, which researchers have identified as the crux of human language [6–9].

Syntax involves a set of rules for combining words into phrases and sentences. Such rules ultimately define *explicit* syntactic relations among words that can be directly mapped into a graph capturing most of the global features of the underlying rules. Such a network-based approach has provided new insights into semantic webs [10–13]. Capturing global syntactic information using a network has been attempted. The global structure of word interactions in short contexts in sentences has been studied [14,15]. Although about 87% of syntactic relationships take place at distances lower than or equal to 2 [16], such early work lacks both a linguistically precise definition of link and fails in capturing the characteristic long-distance correlations of words in sentences [17]. The proportion of incorrect syntactic dependency links captured with a window of length 2 as in Ref. [14] is

$$\epsilon_2 = \frac{(n-1)(1-p_1)+(n-2)(1-p_2)}{2n-3},$$

where n is the length of the sentence and p_1 and p_2 are, respectively, the probability that two words at distances 1 and 2 are syntactically linked. When $n \to \infty$ we have

$$\epsilon_2 = 1 - \frac{p_1 + p_2}{2}.$$

Knowing $p_1 = 0.70$ and $p_2 = 0.17$ [16] we get

$$\epsilon_2 = 0.56.$$

That is, one half of the links are syntactically meaningless. Using a window of length 1 we have

$$\epsilon_1 = \frac{(n-1)(1-p_1)}{n-1}.$$

When $n \to \infty$ we get $\epsilon_1 = 1 - p_1$, which gives $\epsilon_1 = 0.30$, which is still high. A precise definition of syntactic link is thus required. In this paper we study the architecture of syntactic graphs and show that they display small world patterns, scale-free structure, a well-defined hierarchical organization, and assortative mixing [18–20]. Three different European languages will be used. The paper is organized as follows. The three data sets are presented together with a brief definition of the procedure used for building the networks in Sec. II. The key measures used in this study are presented in Sec. III, with the basic results reported in Sec. IV. A comparison between sentence-level patterns and global patterns is presented in Sec. V. A general discussion and summary are given in Sec. VI.

II. THE SYNTACTIC DEPENDENCY NETWORK

The networks that are analyzed here have been defined according to the dependency grammar formalism. Dependency grammar is a family of grammatical formalisms [21–23], which share the assumption that syntactic structure consists of lexical nodes (representing words) and binary relations (dependencies) linking them. This formalism thus naturally defines a network structure. In this approximation, a dependency relation connects a pair of words. Most of the links are directed and the arc goes from the head word to its modifier or vice versa depending on the convention used. Here we assume that links go from the modifier to its head.

*Corresponding author. Email address: rferrer@imim.es

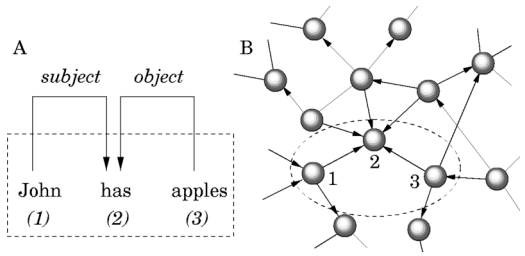

FIG. 1. (a) The syntactic structure of a simple sentence. Here words define the nodes in a graph and the binary relations (arcs) represent syntactic dependencies. Here we assume arcs go from a modifier to its head. The proper noun "John" and the verb "has" are syntactically dependent in the sentence. John is a modifier of the verb has, which is its head. Similarly, the action of has is modified by its object "apples." (b) Mapping the syntactic dependency structure of the sentence in (a) into a global syntactic dependency network.

Head and modifier are primitive concepts in the dependency grammar formalism [Fig. 1(a)]. In some cases, such as coordination, there is no clear direction [24]. Since these cases are rather uncommon, we will assume that links in the data sets used here have a direction and assign an arbitrary direction to the undirected cases. Syntactic relations are thus binary, usually directed and sometimes typed in order to distinguish different kinds of dependencies.

We define a syntactic dependency network as a set of n words $V=\{s_i\}(i=1,\ldots,n)$ and an adjacency matrix $A=\{a_{ij}\}$. s_i can be a modifier word of the head s_j in a sentence if $a_{ij}=1$ ($a_{ij}=0$ otherwise). Here, we assume arcs go from a modifier to its head. The syntactic dependency structure of a sentence can be seen as a subset of all possible syntactic links contained in a global network [Fig. 1(b)]. More precisely, the structure of a sentence is a subgraph (a tree) of the global network that is induced by the words in the sentence [25].

Different measures can be defined on A allowing one to test the presence of certain interesting features such as the small world effect [26] and scale invariance [27]. Such measures can also be used for finding similarities and differences among different networks (see Sec. III).

The common formal property of dependency representations (compared to other syntactic representations) is the lack of explicit encoding for phrases as in the phrase-structure formalism [17] and later developments [28]. Dependency grammar regards phrases as emergent patterns of syntactic dependency interactions. Statistical studies about phrase-structure-based grammars have been performed and reveal that the properties of syntactic constructs map to only a few distributions [3,4], suggesting a reduced set of principles behind syntactic structures.

We studied three global syntactic dependency networks from three European languages: Czech, German, and Romanian. Because of the reduced availability of data, the language set is unintentionally restricted to the Slavic, Germanic, and Italic families. These languages are not intended to be representative of every family. We mention the families these languages belong to in order to show how distant these languages are; probably not enough distant for standard methods in linguistics for finding universals but enough distant for our concerns here. Syntactic dependency networks were built collecting all words and syntactic dependency links appearing in three corpora (a corpus is a collection of sentences). Here $a_{ij}=1$ if an arc from the ith word to the jth word has appeared in a sentence at least once and $a_{ij}=0$ otherwise. Punctuation marks and loops (arcs from a word to itself) were rejected in all three corpora. Sentences with less than two words were rejected.

The corpora analyzed here are the following.

(1) A Czech dependency corpus was annotated by Uhlířová and Králík among others [29,30]. The corpus was compiled at the Czech Language Institute, Prague, within a period of 1970–1985. The corpus contains 562 820 words and 31 701 sentences. Many sentence structures are incomplete in this (i.e., they have less than $n-1$ links, where n is the length of the sentence). The proportion of links provided with regard to the theoretical maximum is about 0.65. The structure of sentences was determined by linguists by hand.

(2) The Romanian corpus was formed by all sample sentences in the Dependency Grammar Annotator website [49]. It contains 21 275 words and 2340 sentences. The syntactic annotation was performed by hand.

(3) The German corpus is The Negra Corpus 1.0. It contains 153 007 words and 10 027 sentences. The formalism used is based on the phrase-structure grammar. Nonetheless, for certain constituents, the head word is indicated. Only the head modifier links between words at the same level of the derivation tree were collected. The syntactic annotation was performed automatically. The proportion of links provided with regard to the theoretical maximum is about 0.16.

The German corpus is the most sparse of them. It is important to notice that while the missing links in the German corpus obey no clear regularity, links in the Czech corpus are mostly function words, specially prepositions, the annotators did not link because they treated them as grammatical markers. The links that are missing are those corresponding to the most connected word types in the remaining corpora.

III. NETWORK PROPERTIES

In order to properly look for syntactic dependency patterns, we need to consider several statistical measures mainly based on the undirected version of the network for simplicity reasons. These measures allow one to categorize networks in terms of the following.

(*a*) *Small world structure*. Two key quantities allow one to characterize the global organization of a complex network. The first is the so called *average path length D*, defined as $D=\langle D_{min}(i,j)\rangle$, where $\langle\cdots\rangle$ is the average operator over all pairs (s_i,s_j) in the network, where $D_{min}(i,j)$ indicates the length of the shortest path between nodes i and j. D was calculated on the largest connected component of the networks. The second measure is C, the so called clustering coefficient, defined as the probability that two vertices (e.g., words) that are neighbors of a given vertex are neighbors of each other. C is defined as $\langle C_i\rangle$ where $\langle\cdots\rangle$ is the average over all vertices and C_i, the clustering coefficient of the ith vertex, is easily defined from the adjacency matrix as

$$C_i = \frac{2}{k_i(k_i-1)} \sum_{j=1}^{n} a_{ij} \left(\sum_{l=j+1} a_{il} a_{jl} \right), \quad (1)$$

where k_i is the degree of the ith vertex. Erdös-Rényi graphs have a binomial degree distribution that can be approximated by a Poissonian distribution [18–20]. Erdös-Rényi graphs with an average degree $\langle k \rangle$ are such that $C_{random} \approx \langle k \rangle / (n-1)$ and the path length follows [31]:

$$D_{random} \approx \frac{\ln n}{\ln \langle k \rangle}. \quad (2)$$

It is said that a network exhibits the small world phenomenon when $D \approx D_{random}$ [26]. The key difference between an Erdös-Rényi graph and a real network is often $C \gg C_{random}$ [18–20].

(*b*) *Heterogeneity*. A different type of characterization of the statistical properties of a complex network is given by the degree distribution $P(k)$. Although the degree distribution of Erdös-Rényi graphs is Poisson, most complex networks are actually characterized by highly heterogeneous distributions: they can be described by a degree distribution $P(k) \sim k^{-\gamma} \phi(k/k_c)$, where $\phi(k/k_c)$ introduces a cutoff at some characteristic scale k_c. The simplest test of scale invariance is thus performed by looking at $P(k)$, the probability that a vertex has degree k, often obeying [18–20]

$$P(k) \sim k^{-\gamma}.$$

The degree distribution is the only statistical measure where link direction will be considered. Therefore, input and output degree will be also analyzed.

(*c*) *Hierarchical organization*. Some scaling properties indicate the presence of hierarchical organization and modularity in complex networks. When studying $C(k)$, i.e., the clustering coefficient as a function of the degree k, certain networks have been shown to behave as [32,33]

$$C(k) \sim k^{-\theta}, \quad (3)$$

with $\theta \approx 1$ [32]. Hierarchical patterns are specially important here, since treelike structures derived from the analysis of sentence structure strongly claim for a hierarchy.

(*d*) *Betweenness centrality*. While many real networks exhibit scaling in their degree distributions, the value of the exponent γ is not universal, the betweenness centrality distribution is less varying [34] although it fails to work as a network classification method [35]. The betweenness centrality of a vertex v, $g(v)$, is a measure of the number of minimum distance paths running through v, which is defined as [34]

$$g(v) = \sum_{i \neq j} \frac{G_v(i,j)}{G(i,j)},$$

where $G_v(i,j)$ is the number of shortest pathways between i and j running through v and $G(i,j) = \Sigma_v G_v(i,j)$. Many real networks obey

$$P(g) \sim g^{-\eta},$$

where $P(g)$ is the proportion of vertices whose betweenness centrality is g. The betweenness centrality was calculated using Brandes' algorithm [36].

e Assortativeness. A network is said to show assortative mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections. A network is said to show disassortative mixing if the highly connected nodes tend to be connected to nodes with few connections. The Pearson correlation coefficient Γ defined in Ref. [37] measures the type of mixing with $\Gamma > 0$ for assortative mixing and $\Gamma < 0$ for disassortative mixing. Such correlation function can be defined as

$$\Gamma = \frac{c \sum_i j_i k_i - \left[c \sum_i \frac{1}{2}(j_i + k_i) \right]^2}{c \sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[c \sum_i \frac{1}{2}(j_i + k_i) \right]^2}, \quad (4)$$

where j_i and k_i are the degrees of the vertices at the ends of the ith edge, with $i = 1, \ldots, m$, $c = 1/m$ and m being the number of edges. Disassortative mixing ($\Gamma < 0$) is shared by Internet, World Wide Web, protein interactions, neural networks, and food webs. In contrast, different kinds of social relationships are assortative ($\Gamma > 0$) [37,38].

IV. RESULTS

The first relevant result of our study is the presence of small world structure in the syntax graph. As shown by our analysis (see Table I for a summary), syntactic networks show $D \approx 3.5$ degrees of separation. The values of D and C are very similar for Czech and Romanian. A certain degree of variation for German can be attributed to the fact that it is the most sparse data set. Thus, D is overestimated and C is underestimated. Nonetheless, all networks have D close to D_{random} which is the hallmark of the small world phenomenon [26]. The fact that $C \gg C_{random}$ indicates (Table I) that the organization of syntactic networks strongly differs from the Erdös-Rényi graphs. Additionally, we have also studied the frequency of short path lengths for the three networks. As shown in Fig. 2, the three distributions are actually very similar, thus suggesting a common pattern of organization. When we compare the observed distributions to the expectation from a random Poissonian graph (indicated by filled triangles), they strongly differ. Although the average value is the same, syntactic networks are much more narrowly distributed. This was earlier observed in the analysis of World Wide Web [39].

The second result concerns the presence of scaling in their degree distributions. The scaling exponents are summarized in Table I. For the undirected graph, we have found that the networks are scale free with $\gamma \approx 2.2$. Additionally, Fig. 3 shows $P(k)$ for input and output degrees (see Table I for the specific values observed). With the exception of the Czech corpus, they display well-defined scale-free distributions. The Czech data set departs from the power law for $k > 10^2$. Thus highly connected words appear underestimated in this

TABLE I. A summary of the basic features exhibited by the three syntactic dependency networks analyzed here. n is the number of vertices of the networks, $\langle k \rangle$ is the average degree, C is the clustering coefficient, and C_{random} is the value of C of an Erdös-Rényi network. D is the average minimum vertex-vertex distance and D_{random} is the value of D for an Erdös-Rényi graph. Γ is the Pearson correlation coefficient. γ, γ_{in}, and γ_{out} are, respectively, the exponents of the undirected degree distribution, input degree distribution, and output degree distribution. η, θ, and ζ are, respectively, the exponents of the betweenness centrality distribution, the clustering vs degree, and the frequency vs degree (estimated within $1<k<10^3$). Two further examples of complex networks are shown. One is a technological graph (a software network analyzed in Ref. [40]) and the second is a biological web: the protein interaction map of yeast [41]. Here *skewed* indicates that the distribution $C(k)$ decays with k but not necessarily following a power law.

	Czech	German	Romanian	Software graph	Proteome[a]
n	33336	6789	5563	1993	1846
$\langle k \rangle$	13.4	4.6	5.1	5.0	2.0
C	0.1	0.02	0.09	0.17	2.2×10^{-2}
C_{random}	4×10^{-4}	6×10^{-6}	9.2×10^{-4}	2×10^{-3}	1.5×10^{-3}
D	3.5	3.8	3.4	4.85	7.14
D_{random}	4	5.7	5.2	4.72	9.0
Γ	−0.06	−0.18	−0.2	−0.08	−0.16
γ	2.29±0.09	2.23±0.02	2.19±0.02	2.85±0.11	$2.5 (k_c \sim 20)$
γ_{in}	1.99±0.01	2.37±0.02	2.20±0.01		
γ_{out}	1.98±0.01	2.09±0.01	2.20±0.01		
η	1.91±0.01	2.10±0.01	2.10±0.01	2.0	2.2
θ	Skewed	Skewed	Skewed	Skewed	1.0
ζ	1.03±0.02	1.18±0.01	1.06±0.02		

[a]Data available from Ref. [50].

case, consistent with the limitations of this corpus discussed in Sec. II. These power laws fully confirm the presence of scaling at all levels of language organization [5].

Complex networks display hierarchical structure [32]. Figure 4 (left column) shows the distribution of clustering coefficients $C(k)$ against degree for the different corpora. We observe skewed distributions of $C(k)$ (which are not power

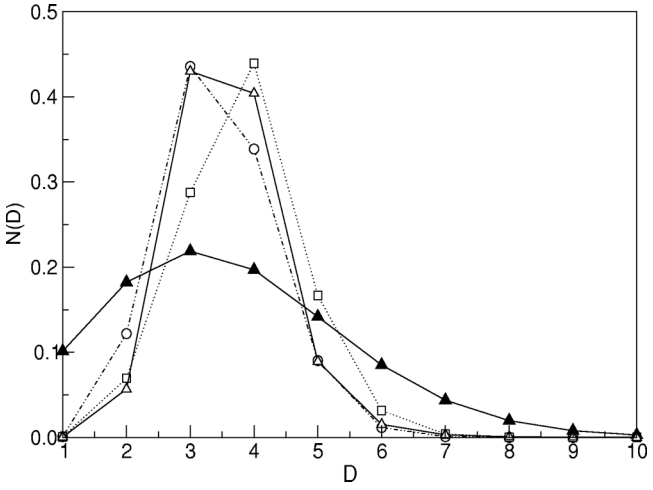

FIG. 2. Shortest path length distributions for the three syntactic networks analyzed here. The symbols correspond to Romanian (circles), Czech (triangles), and German (squares), respectively. The three distributions are peaked around an average distance of $D \approx 3.5$ degrees of separation. The expected distribution for a Poissonian graph is also shown (filled triangles), using the same average distance.

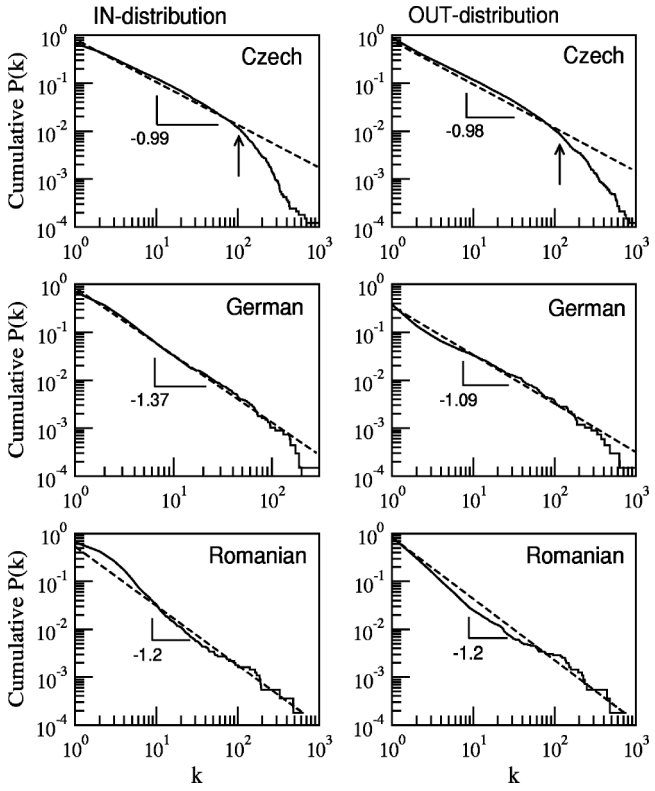

FIG. 3. Left: Cumulative degree distributions for the three corpora. Here the proportion of vertices whose input and output degrees are k is shown. The plots are computed using the cumulative distributions $P_{\geqslant}(k)=\Sigma_{j\geqslant k}P(j)$. The arrows in the plots on top indicate the deviation from the scaling behavior in the Czech corpus (see Sec. IV).

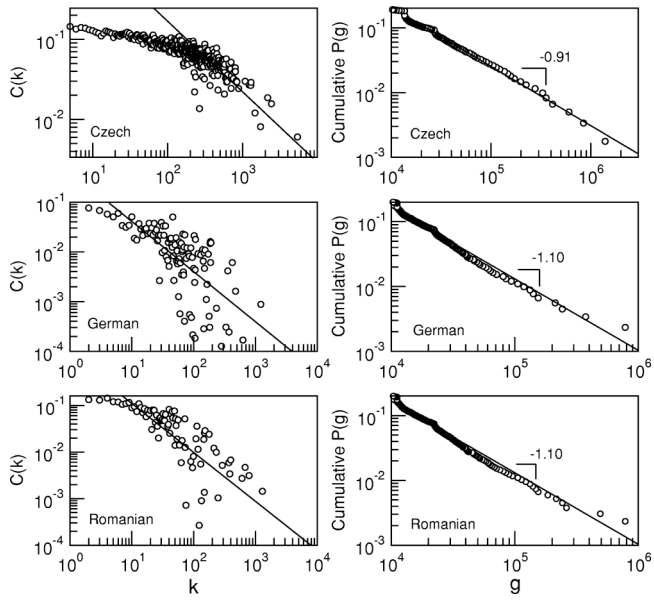

FIG. 4. Left: $C(k)$, the clustering coefficient vs degree k for the three corpora. In all three pictures the scaling relation $C(k) \sim k^{-1}$ is shown for comparison. Right: the corresponding (cumulative) $P(g)$, the proportion of vertices whose betweenness centrality is g.

laws) as in other systems displaying hierarchical organization, such as the World Wide Web (see Fig. 3(c) in Ref. [33]).

In order to measure to what extent word syntactic dependency degree k is related to word frequency f, we calculated the average value of f versus k (5) and found a power distribution of the form

$$f \sim k^{\zeta}, \tag{5}$$

where $\zeta \approx 1$ (Table I) indicates a linear relationship (Fig. 5). The higher values of ζ for German can be attributed to the sparseness of the German corpus. Knowing that Zipf's law states that [2]

$$P(f) \sim f^{-\beta}$$

with typically $\beta \approx 2$, it follows

$$P(k) \sim k^{-\gamma'}$$

with typically $\gamma' \approx 2$ if $\zeta \approx 1$. The estimated γ' is close to the values of γ in Table I.

Highly connected words tend to be not interconnected among them. Since degree and frequency are positively correlated [Eq. (5) and Fig. 5)] one easily concludes, as a visual examination will reveal, that the most connected words are function words (i.e., prepositions, articles, determiners, etc.). Disassortative mixing ($\Gamma < 0$) tells us that function words tend to avoid linking each other. This consistently explains why the Czech corpus has a value of Γ clearly greater than that of the remaining languages. We already mentioned in Sec. II that most of the missing links in the Czech corpus are those involving function words such as prepositions, which are in turn the words responsible for a tendency to avoid links among highly connected words. Γ is thus overestimated in the Czech network.

The scaling exponent γ is somewhat variable, but the scaling exponents obtained for the betweenness centrality measure are more narrowly constrained (Table I). Although again the Czech corpus deviates from the other two (in an expected way), the two other corpora display a remarkable similarity, $P(g)$ distribution with $\eta = 2.1$. Is is worth mentioning that the fits are very accurate and give an exponent that seems to be different from those reported in most complex networks analyzed so far, typically $\eta \in [2.0, 2.2]$ [34]. The behavior of $P(g)$ in Fig. 4 with a domain with scaling with $\eta \approx 2.1$ for German and Romanian suggests a common pattern is shared. The deviation of Cezch from the remaining networks may be explained by its lack of hub words.

The behavior of $C(k)$ (Fig. 4, left) differs from the independence of the vertex degree found in Poisson networks and certain scale-free network models [33]. Such behavior $C(k)$ is also different from Eq. (3) with $\theta = 1$ that is clearly found in synonymy networks and suggested in actor networks [32] and metabolic networks [32]. In contrast, such behavior is similar to that of the World Wide Web and Internet at the autonomous system level [33]. The similar shape of $C(k)$ in the three syntactic dependency networks suggests a common mechanism of hierarchical organization.

Besides word cooccurrence networks and the syntactic dependency networks presented here, other types of linguistic networks have been studied. Networks where nodes are words or concepts and links are semantic relations are known to show $C \gg C_{random}$ with $d \approx d_{random}$ and power distribution of degrees with and exponent $\gamma \in [3, 3.5]$. For Roget's Thesaurus, assortative mixing ($\Gamma = 0.157$) is found [20,10–13]. In contrast, syntactic dependency networks have $\gamma \in [2.11, 2.29]$ and disassortative mixing (Table I), suggesting semantic networks are shaped by radically different factors. Further work, including more precise measures, should be carried out for semantic networks.

V. GLOBAL VERSUS SENTENCE-LEVEL PATTERNS

One may argue that the regularities encountered here are not significant unless it is shown that they are not a trivial

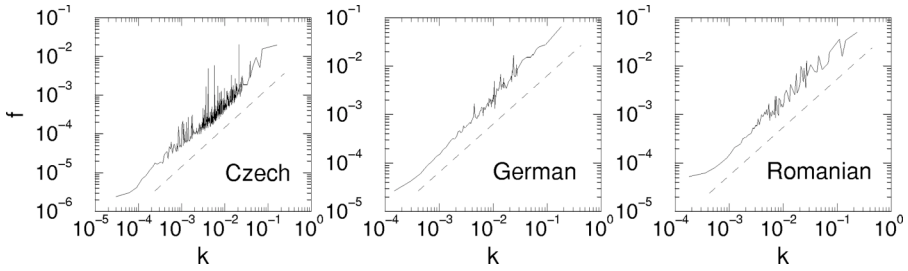

FIG. 5. Average word frequency f of words having degree k. Dashed lines indicate the slope of $f \sim k$, in agreement with real series.

TABLE II. Summary of global vs sentence network traits. d_{global}, C_{global}, and Γ_{global} are, respectively, the normalized average vertex-vertex distance, the clustering coefficient, and the Pearson correlation coefficient of a given global syntactic dependency network. $d_{sentence}$, $C_{sentence}$, and $\Gamma_{sentence}$ are, respectively, the normalized average vertex-vertex distance, the clustering coefficient, and the Pearson correlation coefficient of a given sentence syntactic dependency network. $\langle x \rangle$ stands for the average value of x over all sentence syntactic dependency networks where x is defined.

	Czech	Romanian	German
d_{global}	2.3×10^{-4}	1.3×10^{-3}	1.2×10^{-3}
$\langle d_{sentence} \rangle$	0.88	0.75	0.83
C_{global}	0.1	0.09	0.02
$\langle C_{sentence} \rangle$	0	0	0
Γ_{global}	−0.06	−0.2	−0.18
$\langle \Gamma_{sentence} \rangle$	−0.4	−0.51	−0.64

consequence of some pattern already present in the syntactic structure of isolated sentences. In order to dismiss such possibility, we define d_{global} and $d_{sentence}$ as the normalized vertex-vertex distance of the global dependency networks and a sentence dependency network. The normalized average vertex-vertex distance is defined here as

$$d = \frac{D-1}{D_{max}-1},$$

where $D_{max}=n+1/3$, the maximum distance of a connected network with n nodes [42]. Similarly, we define C_{global} and $C_{sentence}$ for the clustering coefficient and Γ_{global} and $\Gamma_{sentence}$ for the Pearson correlation coefficient. The clustering coefficient of whatever syntactic dependency structure is $C_{sentence}=0$, since the syntactic dependency structure is defined with no cycles [21]. We find $C_{global} \gg C_{sentence}$ and $d_{global} \ll d_{sentence}$ and d_{global} significantly smaller than $d_{sentence}$ (Table II). $\Gamma_{sentence}$ is clearly different than Γ_{global}, although disassortative mixing is found in both cases.

Besides, one may think that the global degree distribution is scale-free because the degree distribution of the syntactic dependency structure of a sentence is already scale-free. $P_{sentence}(k)$, the probability that the degree of a word in a sentence is k is not a power function of k (Fig. 6). Actually, the data point suggests an exponential fit. To sum up, we conclude that scaling in $P(k)$, small world with significantly high C, and the global value of Γ are features emerging at the macroscopic scale. The global patterns discussed above are emergent features that show up at the global level.

VI. DISCUSSION

We have presented a study of the statistical patterns of organization displayed by three different corpora in this paper. The study reveals that, as it occurs at other levels of language organization [10–13], scaling is widespread. The analysis shows that syntax is a small world and suggests other potentially broad patterns for languages on Earth. Such

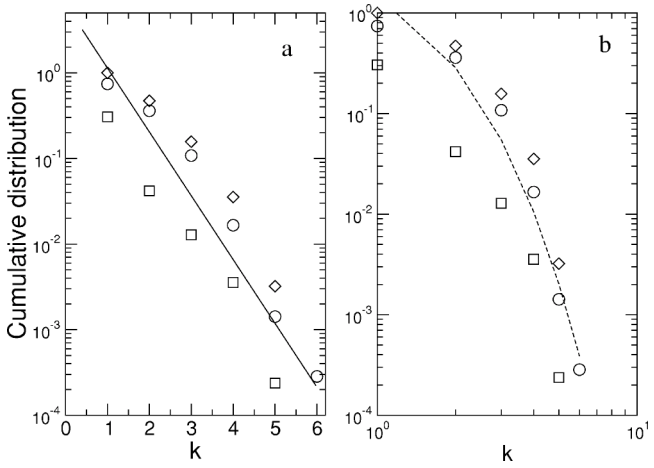

FIG. 6. Cumulative $P_{sentence}(k)$ for Czech (circles), German (squares), and Romanian (diamonds). Here linear-log (a) and log-log (b) plots have been used, indicating an exponential-like decay. $P_{sentence}(k)$ is the probability that a word has degree k in the syntactic dependency structure of a sentence. Notice that $P_{\geqslant}(1)$ is less than 1 for Czech and German since the sentence dependency trees are not complete. If $P_{sentence}$ was a power function, a straight line should appear in log-log scale. The German corpus is so sparse that its appearance is dubious. Statistics are shown for L^*, the typical sentence length. We have $L^*=12$ for Czech and German and $L^*=6$ for Romanian. The average value of $P_{sentence}(k)$ for all sentence lengths is not used since it can be misleading as [16] shows in a similar context.

patterns rely on precise measures and have been shown to be rather homogeneous.

Understanding the origins of syntax implies understanding what is essential in human language. Recent studies have explored this question by using mathematical models inspired by evolutionary dynamics [43–45]. However, the study of the origins of language is usually dissociated from the quantitative analysis of real syntactic structures. The statistical pattern reported here could serve as validation of existent formal approaches to the origins of syntax. What is reported here is specially suitable for evolutionary approaches to the origins of language, since they reduce syntax to word pairwise relationships.

Linguists can decide not to consider certain word types as vertices in the syntactic dependency structure. For instance, annotators in the Czech corpus decided that prepositions are not vertices. This way, we have seen that different statistical regularities are distorted, e.g., disassortative mixing almost disappears and degree distributions are truncated with regard to the remaining corpora. If the degree distribution is truncated, describing degree distributions requires more complex functions. If simplicity is a desirable property, syntactic descriptions should consider prepositions and similar word types as words in the strict sense. Annotators should be aware of the consequences of their decision about the local structure of sentences with regard to global statistical patterns.

Syntactic dependency networks do not imply recursion, which is regarded as a crucial trait of the language faculty [6]. Nonetheless, different nontrivial traits that recursion needs have been quantified:

(1) Disassortative mixing tells us that labor is divided in human language. Linking words tend to avoid connections among them.

(2) Hierarchical organization tells us that syntactic dependency networks not only define the syntactically correct links (if certain context freedom is assumed) but also a top-down hierarchical organization that is the basis of phrase-structure formalisms [46].

(3) Small worldness is a necessary condition for recursion. If mental navigation [13] in the syntactic dependency structure cannot be performed reasonably fast, recursion cannot take place. In this context, pressures for fast vocal communication are known to exist [47,48].

Regardless of the heterogeneity of the annotation criteria, common patterns have appeared, suggesting interesting prospects for future deeper and broader studies. The present work is a starting point for finding linguistic universals from the point of view of complex networks. The patterns presented here are candidates for linguistic universals. More empirical and theoretical work is needed to establish such syntactic dependency universals.

ACKNOWLEDGMENTS

We thank Ludmila Uhlířová for providing us with the opportunity to analyze the Czech corpus for the present study. R.F.C. thanks a grant from the Generalitat de Catalunya (Grant No. FI/2000-00393). This work has been also supported by Grant No. BFM 2001-2154 and by the Santa Fe Institute (RVS).

[1] D. Crystal, *The Cambridge Encyclopedia of Language* (Cambridge University Press, Cambridge, UK, 1997).
[2] G. K. Zipf, *Human Behaviour and the Principle of Least Effort. An introduction to Human Ecology*, 1st ed. (Addison-Wesley, Cambridge, MA, 1949); reprinted by (Hafner, New York, 1972).
[3] R. Köhler, J. Quantitative Linguistics **6**, 46 (1999).
[4] R. Köhler and G. Altmann, J. Quantitative Linguistics **7**, 189 (2000).
[5] L. Hřebíček, *Quantitative Linguistics*, edited by G. Altmann, R. Köhler, and B. Rieger (Wissenschaftlicher Verlag, Trier, 1995), Vol. 56.
[6] M. D. Hauser, N. Chomsky, and W. T. Fitch, Science **298**, 1569 (2002).
[7] T. W. Deacon, *The Symbolic Species: The Co-evolution of Language and the Brain* (Norton, New York, 1997).
[8] M. Donald, *Origins of the modern mind* (Cambridge University Press, Cambridge, MA, 1991).
[9] M. Donald, in *Approaches to the Evolution of Language: Social and Cognitive Bases*, edited by J. R. Hurford, M. Studdert-Kennedy, and C. Knight (Cambridge University Press, Cambridge, 1998), pp. 44–67.
[10] A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta, Phys. Rev. E **65**, 065102 (2002).
[11] M. Sigman and G. A. Cecchi, Proc. Natl. Acad. Sci. U.S.A. **99**, 1742 (2002).
[12] M. Steyvers and J. Tenenbaum, e-print cond-mat/0110012.
[13] O. Kinouchi, A. S. Martinez, G. F. Lima, G. M. Lourenço, and S. Risau-Gusman, Physica A **315**, 665 (2002).
[14] R. Ferrer i Cancho and R. V. Solé, Proc. R. Soc. London, Ser. B **268**, 2261 (2001).
[15] S. N. Dorogovtsev and J. F. F. Mendes, Proc. R. Soc. London, Ser. B **268**, 2603 (2001).
[16] R. Ferrer i Cancho (unpublished).
[17] N. Chomsky, *Syntactic Structures* (Mouton, S-Gravenhage, 1957).
[18] A.-L. Barabási and R. Albert, Rev. Mod. Phys. **74**, 47 (2002).
[19] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[20] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[21] I. Melčuk, *Dependency Syntax: Theory and Practice* (SUNY, New York, 1988).
[22] R. Hudson, *Word Grammar* (Blackwell, Oxford, 1984).
[23] D. Sleator and D. Temperley, Carnegie Mellon University Technical Report No. CMU-CS-91-196, 1991 (unpublished).
[24] I. Melčuk, in *International Encyclopedia of the Social and Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Pergamon, Oxford, 2002), pp. 8336–8344.
[25] B. Bollobás, *Modern Graph Theory*, Graduate Texts in Mathematics Vol. 184 (Springer, New York, 1998).
[26] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[27] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[28] J. Uriagereka, *Rhyme and Reason. An Introduction to Minimalist Syntax* (MIT Press, Cambridge, MA, 1998).
[29] L. Uhlírova, I. Nebeská, and J. Králík, in *COLING 82, Proceedings of the Ninth International Conference on Computational Linguistics, Prague*, edited by J. Horecký (North-Holland, Amsterdam, 1982), pp. 391–396.
[30] M. Těšitelová, *Kvantitativni Charakteristiky Současné Češtini [Quantitative Characteristics of Present-day Czech]* (Academia, Praha, 1985), p. 249s.
[31] M. E. J. Newman, J. Stat. Phys. **101**, 819 (2000).
[32] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, Science **297**, 1551 (2002).
[33] E. Ravasz and A.-L. Barabási, Phys. Rev. E **67**, 026112 (2002).
[34] K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim, Proc. Natl. Acad. Sci. U.S.A. **99**, 12583 (2002).
[35] M. Barthélemy, Phys. Rev. Lett. **91**, 189803 (2003).
[36] U. Brandes, J. Math. Sociol. **25**, 163 (2001).
[37] M. E. J. Newman, Phys. Rev. Lett. **89**, 208701 (2002).
[38] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).
[39] L. A. Adamic, *Proceedings of the ECDL'99 Conference, LNCS 1696* (Springer, Berlin, 1999) pp. 443–452.
[40] S. Valverde, R. Ferrer i Cancho, and R. V. Solé, Europhys. Lett. **60**, 512 (2002).
[41] H. Jeong, S. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).
[42] R. Ferrer i Cancho and R. V. Solé, in *Statistical Mechanics of*

Complex Networks, edited by R. Pastor-Satorras, J. M. Rubi, and A. Diaz-Guilear, Lecture Notes in Physics Vol. 625 (Springer, Berlin, 2003), pp. 114–125.
[43] M. A. Nowak and D. C. Krakauer, Proc. Natl. Acad. Sci. U.S.A. **96**, 8028 (1999).
[44] M. A. Nowak, J. B. Plotkin, and V. A. Jansen, Nature (London) **404**, 495 (2000).
[45] M. A. Nowak, Philos. Trans. R. Soc. London, Ser. B **355**, 1615 (2000).
[46] D. Bickerton, *Language and Species* (Chicago University Press, 1990).
[47] J. A. Hawkins, in *Innateness and Function in Language Universals*, edited by J. A. Hawkins and M. Gell-Mann (Addison-Wesley, Redwood, CA, 1992), pp. 87–120.
[48] P. Lieberman, *Uniquely Human: The Evolution of Speech, Thought and Selfless Behavior* (Harvard University Press, Cambridge, MA, 1991).
[49] See http://phobos.cs.unibuc.ro/roric/DGA/dga.html
[50] http://www.nd.edu/~networks/database/protein/bo.dat.gz