Ferrer i Cancho, Ramon/ Mehler, Alexander/ Pustylnikov, Olga/ Díaz-Guilera, Albert (2007):
Correlations in the organization of large-scale syntactic dependency networks. Proceedings
of the workshop "TextGraphs-2: Graph-based Methods for Natural Language Processing" at the
Annual Conference of the North American Chapter of the Association for Computational
Linguistics (NAACL-HLT 2007), Rochester, New York.

Correlations in the organization of large-scale syntactic dependency networks

Ramon Ferrer i Cancho
Departament de Física Fonamental
Universitat de Barcelona
Martí i Franquès 1, 08028 Barcelona, Spain.
ramon.ferrericancho@ub.edu

Alexander Mehler
Department of Computational Linguistics and Text Technology
Bielefeld University
D-33615 Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Olga Pustylnikov
Department of Computational Linguistics and Text Technology
Bielefeld University
D-33615 Bielefeld, Germany
Olga.Pustylnikov.@uni-bielefeld.de

Albert Díaz-Guilera
Departament de Física Fonamental
Universitat de Barcelona
Martí i Franquès 1, 08028 Barcelona, Spain.
albert.diaz@ub.edu

Abstract

We study the correlations in the connectivity patterns of large scale syntactic dependency networks. These networks are induced from treebanks: their vertices denote word forms which occur as nuclei of dependency trees. Their edges connect pairs of vertices if at least two instance nuclei of these vertices are linked in the dependency structure of a sentence. We examine the syntactic dependency networks of seven languages. In all these cases, we consistently obtain three findings. Firstly, clustering, i.e., the probability that two vertices which are linked to a common vertex are linked on their part, is much higher than expected by chance. Secondly, the mean clustering of vertices decreases with their degree — this finding suggests the presence of a hierarchical network organization. Thirdly, the mean degree of the nearest neighbors of a vertex x tends to decrease as the degree of x grows — this finding indicates disassortative mixing in the sense that links tend to connect vertices of dissimilar degrees. Our results indicate the existence of common patterns in the large scale organization of syntactic dependency networks.

1 Introduction

During the last decade, the study of the statistical properties of networks as different as technical, biological and social networks has grown tremendously. See (Barabási and Albert, 2002; Dorogovtsev and Mendes, 2002; Newman, 2003) for a review. Among them many kinds of linguistic networks have been studied: e.g., free word association networks (Steyvers and Tenenbaum, 2005), syllable networks (Soares et al., 2005), thesaurus networks (Sigman and Cecchi, 2002), and document networks (Mehler, 2006). See (Mehler, 2007a) for a review of linguistic network studies. Here we focus on the so called *global syntactic dependency networks* (GSDN) (Ferrer i Cancho et al., 2004; Ferrer i Cancho, 2005). A GSDN is induced from a dependency treebank in two steps:

1. The vertices of the network are obtained from the word forms appearing as nuclei in the input treebank and from punctuation marks as far as they have been annotated and mapped onto dependency trees. The notion of a nucleus is adapted from Lucien Tesnière: a nucleus is a node of a dependency tree. Note that multipart nuclei may also occur. We use the term *type* in order to denote word forms and punctuation marks. The reason that we induce vertices from types, but not from lexemes, is that not all corpora are lemmatized. Thus, the type level is the *least common denominator* which allows comparing the different networks. Note also that a systematization of the corpora with respect to the inclusion of punctuation marks is needed.

2. Two vertices (i.e. types) of a GSDN are connected if there is at least one dependency tree in which their corresponding instance nuclei are linked. When it comes to applying the apparatus of complex network theory, the arc direction is generally disregarded (Newman, 2003). Thus, GSDNs are *simple undirected* graphs without loops or multiple edges.

The attribute 'global' distinguishes *macroscopic* syntactic dependency networks from their *microscopic* counterparts in the form of syntactic dependency structures of single sentences. The latter are the usual object of dependency grammars and related formalisms. The goal of this article is to shed light on the large-scale organization of syntactic dependency structures. In terms of theoretical linguistics, we aim to determine the statistical properties that are common to all languages (if they exist), the ones that are not and to explain our findings. To achieve this goal, we must overcome the limits of many studies of linguistic networks. Firstly, by using GSDNs we intend to solve the problems of co-occurrence networks in which words are linked if they (a) are adjacent, (b) co-occur within a short window (Ferrer i Cancho and Solé, 2001; Milo et al., 2004; Antiqueira et al., 2006; Masucci and Rodgers, 2006) or (c) appear in the same sentence (Caldeira et al., 2006). This approach is problematic: with a couple of exceptions (Bordag et al., 2003; Ferrer i Cancho and Solé, 2001), no attempt is made to filter out statistically insignificant co-occurrences. Unfortunately the filter used in (Ferrer i Cancho and Solé, 2001) is not well-defined because it does not consider fluctuations of the frequencies of word co-occurrences. (Bordag et al., 2003) implement a collocation measure based on the Poisson distribution and, thus, induce *collocation* instead of *co-occurrence* networks. However, the notion of a sentence window and related notions are problematic as the probability that two words depend syntactically decays exponentially with the number of intermediate words (Ferrer i Cancho, 2004). Further, (Ferrer i Cancho et al., 2004) shows that the proportion of syntactically wrong links captured from a sentence by linking adjacent words is about 0.3 while this proportion is about 0.5 when linking a word to its 1st and 2nd neighbors. Thus, dependency treebanks offer connections between words that are *linguistically* precise according to a dependency grammar formalism. Secondly, the majority of linguistic network studies is performed on English only — with some exceptions (Soares et al., 2005; Ferrer i Cancho et al., 2004; Mehler, 2006). Concerning GSDNs, (Ferrer i Cancho et al., 2004) considers three languages but the syntactic dependency information of sentences is systematically incomplete in two of them. Here we aim to use complete treebanks and analyze more (i.e. seven) languages so that we can obtain stronger conclusions about the common statistical patterns of GSDNs than in (Ferrer i Cancho et al., 2004).

Therefore, this article is about statistical regularities of the organization of GSDNs. These networks are analyzed with the help of complex network theory and, thus by means of quantitative graph theory. We hypothesize that GSDNs are homogeneous in terms of their network characteristics while they differ from *non-syntactic* networks. The long-term objective to analyze such distinctive features is to explore *quality criteria* of dependency treebanks which allow separating high quality annotations from erroneous ones.

The remainder of this article is organized as follows: Section 2 introduces the statistical measures that will be used for studying GSDNs of seven languages. Section 3 presents the treebanks and their unified representations from which we induce these networks. Section 4 shows the results and Section 5 discusses them.

2 The statistical measures

Two essential properties of a network are N, the number of vertices (i.e. the number of types), and $\bar{k}$ the mean vertex degree (Barabási and Albert, 2002). The literature about distinctive indices and distributions of complex networks is huge. Here we focus on correlations in the network structure (Serrano et al., 2006). The reason is that correlation analysis provides a deeper understanding of network organization compared to classical aggregative "small-world" indices. For instance, two networks may have the same degree distribution (whose similarity is measured by the exponent of power laws fitted to them) while they differ in the degree correlation of the vertices forming a link. Correlation analysis is performed as follows: We define $p(k)$ as the proportion of vertices with degree k. Here we study three measures of correlation (Serrano et al., 2006):

- $\bar{k}_{nn}(k)$ is the average degree of the nearest neighbors of the vertices with degree k (Pastor-Satorras et al., 2001). If $\bar{k}_{nn}(k)$ tends to grow as k grows the network is said to exhibit assortative mixing. In this case, edges tend to connect vertices of similar degree. If in contrast to this $\bar{k}_{nn}(k)$ tends to shrink as k grows, the network is said to exhibit disassortative mixing. In this case, edges tend to connect vertices of dissimilar degree. If there are no correlations, then $\bar{k}_{nn}(k) = \kappa$ with $\kappa = \langle k^2 \rangle / \langle k \rangle$; $\langle k \rangle = \bar{k}$ is the 1st and $\langle k^2 \rangle$ the 2nd moment of the degree distribution, namely

$$\langle k \rangle = \sum_{k=1}^{N-1} kp(k) \tag{1}$$

$$\left\langle k^2 \right\rangle = \sum_{k=1}^{N-1} k^2 p(k). \tag{2}$$

  In order to enable comparisons of different networks, $\bar{k}_{nn}(k)$ is normalized using κ and replaced by $\bar{k}_{nn}(k)/\kappa$.

- $\bar{c}(k)$ is the mean clustering coefficient of vertices of degree k. The clustering coefficient of a vertex is defined as the proportion of pairs of adjacent vertices (u, v) such that u and v are linked.

- $\bar{c}$ is the mean clustering coefficient defined as

$$\bar{c} = \sum_{k=1}^{N-1} p(k)\bar{c}(k). \tag{3}$$

  In order to test the significance of $\bar{c}$, we calculate $\bar{c}_{binom} = \bar{k}/(N-1)$, the expected clustering coefficient in a control binomial graph. In a binomial graph, two vertices are linked with probability p. $p = \bar{k}/(N-1)$ is chosen so that the expected number of links of the binomial graphs is $n\bar{k}/2$ as in the original network.

Assortative mixing is known to be characteristic for *social-semiotic*, but not for *technical* networks (Newman, 2003). Recently, (Mehler, 2006) has shown that this characteristic varies a lot for different document networks and thus allows distinguishing linguistic networks which are homogeneously called 'small-worlds'. We have excluded on purpose the Pearson correlation coefficient of the degrees at the endpoints of edges that has been used in previous studies (Ferrer i Cancho et al., 2004) due to the statistical problems that this measure has in large networks with power degree distributions (Serrano et al., 2006).

3 The treebanks

We analyze seven treebanks each from a different language. Their features are summarized in Table 1. A comprehensive description of these and related banks is given by (Kakkonen, 2005). As explained by Kakkonen, one generally faces the problem of the heterogeneity not only of the annotation schemes, but also of the serialization formats used by them. Thus, we unified the various formats in order to get a single interface to the analysis of syntactic dependency networks derived thereof. Although there exists a representation format for syntactic annotations (i.e. TIGER-XML — cf. (Mengel and Lezius, 2000)) we decided to use the Graph eXchange Language (GXL) in order to solve the heterogeneity problem. The GXL has been proposed as a uniform format for data interchange (Holt et al., 2006). It allows representing attributed, directed, undirected, mixed, ordered, hierarchical graphs as well as hypergraphs. Its application-dependent attribution model concerns vertices, edges and graphs. Because of its expressiveness it was utilized in modeling constituency structures (Pustylnikov, 2006) as well as nonlinear document structures (Mehler, 2007b). We utilize it to map syntactic dependency structures.

Our GXL binding is schematically explained as follows: corpora are mapped onto graphs which serialize graph models of sentence-related dependency structures. Each of these structures is mapped as a forest whose directed edges are mapped by means of the GXL's edge model. This model preserves the orientation of the input dependency relations. Figure 1 visualizes a sample dependency tree of the Slovene dependency treebank (Džeroski et al., 2006).

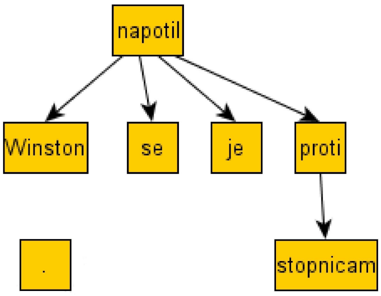

Figure 1: Visualization of a sample sentence of the Slovene dependency treebank (Džeroski et al., 2006) based on its reconstruction in terms of the GXL.

4 Results

A summary of the network measures obtained on the seven corpora is shown in Table 2. We find that $\bar{c} \gg \bar{c}_{binom}$ indicating a clear tendency of vertices connected to be connected if they are linked to the same vertex.

Since the Italian and the Romanian corpus are Romanic languages and the size of their networks is similar, they are paired in the figures. Figure 2 shows that the clustering $\bar{c}(k)$ decreases as k increases. Figure 3 shows that $\bar{k}_{nn}(k)$ decreases as k increases, indicating the presence of disassortative mixing when forming links, i.e. links tend to combine vertices of dissimilar degrees. For sufficiently large k the curves suggest a power-law behavior, i.e. $\bar{k}_{nn}(k) \sim k^{-\eta}$.

5 Discussion

We have found that the behavior of $\bar{k}_{nn}(k)$ suggests $\bar{k}_{nn}(k) \sim k^{-\eta}$ for sufficiently large k. A power-law behavior has been found in technical systems (Serrano et al., 2006). In a linguistic context, a power-law like behavior with two regimes has been found in the word adjacency network examined in (Masucci and Rodgers, 2006). A decreasing $\bar{k}_{nn}(k)$ for growing k (an indicator of dissortative mixing) has been found in biological and social systems (Serrano et al., 2006). A decreasing $\bar{c}(k)$ for growing k has been found in many non-linguistic systems (e.g. the Internet map at the autonomous system level), and also in a preliminary study of Czech and German syntactic dependency networks (Ferrer i Cancho et al., 2004). (Ravasz and Barabási, 2003) suggest that this behavior indicates the existence of a hierarchical network organization (Ravasz and Barabási, 2003). In our case this may indicate the existence of a core vocabulary surrounded by more and more special vocabularies. This observation is in accordance with a multipart organization of the rank frequency distribution of the lexical units involved. But this stratification is not simply due to the words' collocation patterns, but to their behavior in syntactic dependency structures. We have also found that $\bar{c} \gg \bar{c}_{binom}$, which is a common feature of non-linguistic (Newman, 2003) and linguistic networks (Mehler, 2007a) and, thus, is not very informative.

In sum, we have seen that GSDNs follow a common pattern of statistical correlations regardless of the heterogeneity of the languages and annotation criteria used. This suggests that the structure of GSDNs may originate from language independent principles. Since the correlational properties of GSDNs are not unique to these networks, our findings suggest that these principles may also be common to certain non-linguistic systems. Thus, in order to make GSDNs distinguishable in terms of their characteristics, finding more expressive network coeffi-

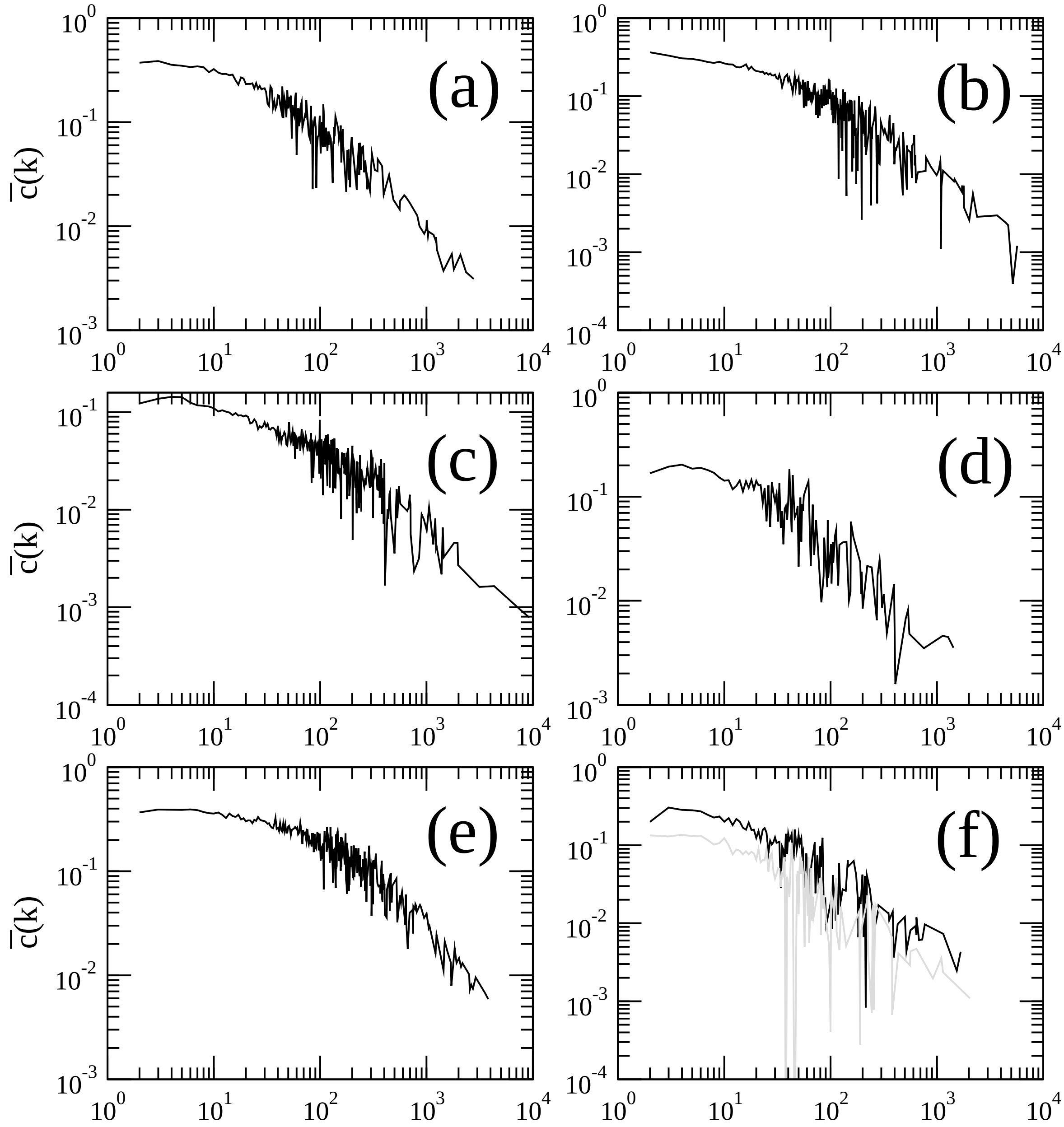

Figure 2: $\bar{c}(k)$, the mean clustering coefficient of vertices of degree k. (a) Danish, (b) Dutch, (c) Russian, (d) Slovene, (e) Swedish and (f) Italian (black) and Romanian (gray).

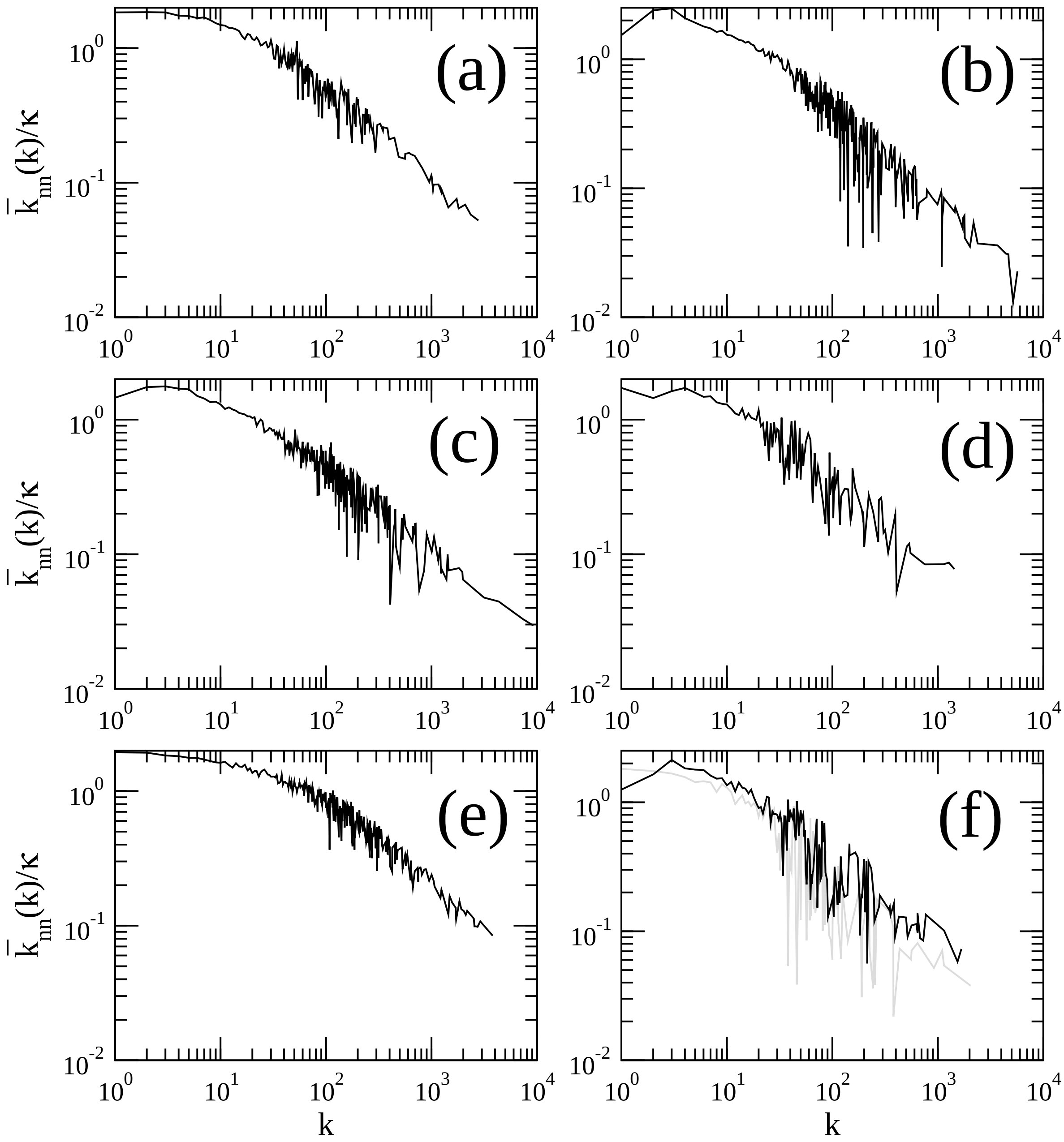

Figure 3: $\bar{k}_{nn}(k)/\kappa$, the normalized mean degree of the nearest neighbors of vertices of degree k. (a) Danish, (b) Dutch, (c) Russian, (d) Slovene, (e) Swedish and (f) Italian (black) and Romanian (gray).

Treebank	Language	Size (#nuclei)	Marks included	Reference
Alpino Treebank v. 1.2	Dutch	195.069	yes	(van der Beek et al., 2002)
Danish Dependency Treebank v. 1.0	Danish	100.008	yes	(Kromann, 2003)
Sample of sentences of the Dependency Grammar Annotator	Romanian	36.150	no	`http://www.phobos.ro/roric/DGA/dga.html`
Russian National Corpus	Russian	253.734	no	(Boguslavsky et al., 2002)
A sample of the Slovene Dependency Treebank v. 0.4	Slovene	36.554	yes	(Džeroski et al., 2006)
Talkbanken05 v. 1.1	Swedish	342.170	yes	(Nivre et al., 2006)
Turin University Treebank v. 0.1	Italian	44.721	no	(Bosco et al., 2000)

Table 1: Summary of the features of the treebanks used in this study. Besides the name, language and version of the corpus we indicate its size in terms of the number of nuclei tokens in the treebank. We also indicate if punctuation marks are treated as vertices of the syntactic structure of sentencess or not.

Language	N	$\bar{k}$	$\bar{c}$	$\bar{c}_{\mathrm{binom}}$
Alpino Treebank v. 1.2	28491	8.1	0.24	0.00028
Danish Dependency Treebank v. 1.0	19136	5.7	0.20	0.00030
Dependency Grammar Annotator	8867	5.3	0.093	0.00060
Russian National Corpus	58285	6.1	0.088	0.00010
Slovene Dependency Treebank v. 0.4	8354	5.3	0.12	0.00064
Talkbanken05 v. 1.1	25037	10.5	0.27	0.00042
Turin University Treebank v. 0.1	8001	6.9	0.18	0.00086

Table 2: Summary of the properties of the GSDNs analyzed. N is the number of vertices, $\bar{k}$ is the mean degree, $\bar{c}$ is the mean clustering coefficient, $\bar{c}_{binom}$ is the clustering coefficient of the control binomial graph.

cients is needed. A possible track could be considering the weight of a link, which is known to provide a more accurate description of the architecture of complex networks (Barrat et al., 2004).

Acknowledgement

We are grateful to Tomaž Erjavec for the opportunity to analyze a sample of the Slovene Dependency Treebank (`http://nl.ijs.si/sdt/`). ADG and RFC's work was funded by the projects FIS2006-13321-C02 and BFM2003-08258-C02-02 of the Spanish Ministry of Education and Science. AM and OP's work is supported by the SFB 673 *Alignment in Communication* (`http://ariadne.coli.uni-bielefeld.de/sfb/`) funded by the German Research Foundation (DFG).

References

Lucas Antiqueira, Maria das Gracas V. Nunes, Osvaldo N. Oliveira, and Luciano da F. Costa. 2006. Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820.

Albert-László Barabási and Réka Albert. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. 2004. The architecture of complex weighted networks. In *Proc. Nat. Acad. Sci. USA*, volume 101, pages 3747–3752.

Igor Boguslavsky, Ivan Chardin, Svetlana Grigorieva, Nikolai Grigoriev, Leonid Iomdin, Leonid Kreidlin, and Nadezhda Frid. 2002. Development of a dependency treebank for russian and its possible applications in NLP. In *Proc. of LREC 2002*.

Stefan Bordag, Gerhard Heyer, and Uwe Quasthoff. 2003. Small worlds of concepts and other principles of semantic search. In *Proc. of the Second International Workshop on Innovative Internet Computing Systems (IICS '03)*.

Cristina Bosco, Vincenzo Lombardo, Daniela Vassallo, and Lesmo Lesmo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proc. of LREC 2000*.

Silvia Maria Gomes Caldeira, Thierry Petit Lobão, Roberto Fernandes Silva Andrade, Alexis Neme, and J. G. Vivas Miranda. 2006. The network of concepts in written texts. *European Physical Journal B*, 49:523–529.

Serguei N. Dorogovtsev and Jose Fernando Ferreira Mendes. 2002. Evolution of random networks. *Adv. Phys.*, 51:1079–1187.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtský, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of LREC 2006*.

Ramon Ferrer i Cancho and Ricard V. Solé. 2001. The small-world of human language. *Proc. R. Soc. Lond. B*, 268:2261–2266.

Ramon Ferrer i Cancho, Ricard V. Solé, and Reinhard Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69:051915.

Ramon Ferrer i Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135.

Ramon Ferrer i Cancho. 2005. The structure of syntactic dependency networks from recent advances in the study of linguistic networks. In V. Levickij and G. Altmann, editors, *The problems in quantitative linguistics*, pages 60–75. Ruta, Chernivtsi.

Richard C. Holt, Andy Schürr, Susan Elliott Sim, and Andreas Winter. 2006. GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170.

Tuomo Kakkonen. 2005. Dependency treebanks: methods, annotation schemes and tools. In *Proc. of NODALIDA 2005*, pages 94–104, Joensuu, Finland.

Matthias T. Kromann. 2003. The Danish dependency treebank and the underlying linguistic theory. In Joakim Nivre and Erhard Hinrichs, editors, *Proc. of TLT 2003*. Växjö University Press.

Adolfo Paolo Masucci and Geoff J. Rodgers. 2006. Network properties of written human language. *Physical Review E*, 74:026102.

Alexander Mehler. 2006. Text linkage in the wiki medium – a comparative study. In *Proc. of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources*, pages 1–8.

Alexander Mehler. 2007a. Large text networks as an object of corpus linguistic studies. In A. Lüdeling and M. Kytö, editors, *Corpus linguistics. An international handbook of the science of language and society*. de Gruyter, Berlin/New York.

Alexander Mehler. 2007b. Structure formation in the web. A graph-theoretical model of hypertext types. In A. Witt and D. Metzing, editors, *Linguistic Modeling of Information and Markup Languages*. Springer, Dordrecht.

Andreas Mengel and Wolfgang Lezius. 2000. An XML-based representation format for syntactically annotated corpora. In *Proc. of LREC 2000*.

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. 2004. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542.

Mark E. J. Newman. 2003. The structure and function of complex networks. *SIAM Review*, 45:167–256.

Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proc. of LREC 2006*.

Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vesipignani. 2001. Dynamical and correlation properties of the internet. *Physical Review Letters*, 87(25):268701.

Olga Pustylnikov. 2006. How much information is provided by text structure? Automatic text classification using structural features (in German). Master thesis, University of Bielefeld, Germany.

Erzsébet Ravasz and Albert-László Barabási. 2003. Hierarchical organization in complex networks. *Phys. Rev. E*, 67:026112.

M. Ángeles Serrano, Marian Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. 2006. Correlations in complex networks. In G. Caldarelli and A. Vespignani, editors, *Structure and Dynamics of Complex Networks, From Information Technology to Finance and Natural Science*, chapter 1. World Scientific.

Mariano Sigman and Guillermo A. Cecchi. 2002. Global organization of the WordNet lexicon. In *Proc. Natl. Acad. Sci. USA*, volume 99, pages 1742–1747.

Márcio Medeiros Soares, Gilberto Corso, and Liacir dos Santos Lucena. 2005. The network of syllables in Portuguese. *Physica A*, 355(2-4):678–684.

Mark Steyvers and Josh Tenenbaum. 2005. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Leonoor van der Beek, Gosse Bouma, Robert Malouf, and Gertjan van Noord. 2002. The Alpino dependency treebank. In *Proc. of the Conf. on Computational Linguistics in the Netherlands (CLIN '02)*.