A LETTERS JOURNAL EXPLORING
THE FRONTIERS OF PHYSICS



# Can syntactic networks indicate morphological complexity of a language?

A LETTERS JOURNAL EXPLORING
THE FRONTIERS OF PHYSICS

January 2011

EPL, **93** (2011) 28005
doi: 10.1209/0295-5075/93/28005

www.epljournal.org

# Can syntactic networks indicate morphological complexity of a language?

HAITAO LIU[1(a)] and CHUNSHAN XU[2]

[1] *School of International Studies, Zhejiang University - CHN-310058 Hangzhou, China*
[2] *Institute of Applied Linguistics, Communication University of China - CHN-100024 Beijing, China*

received 19 October 2010; accepted in final form 10 January 2011
published online 8 February 2011

PACS 89.75.Da – Systems obeying scaling laws
PACS 89.75.Fb – Structures and organization in complex systems
PACS 89.75.Hc – Networks and genealogical trees

**Abstract** – In this study, the complex-network approaches are employed to investigate the word form networks and the lemma networks extracted from dependency syntactic treebanks of fifteen different languages. The results show that it is possible to classify human languages by means of the main parameters of complex networks. The complex-network approaches can obtain language classifications as precise as achieved by contemporary word order typology. Clustering experiments point to the fact that the difference between the word form networks and the lemma networks can make for a better classification of languages. In short, the dependency syntactic networks can reflect morphological variation degrees and morphological complexity.



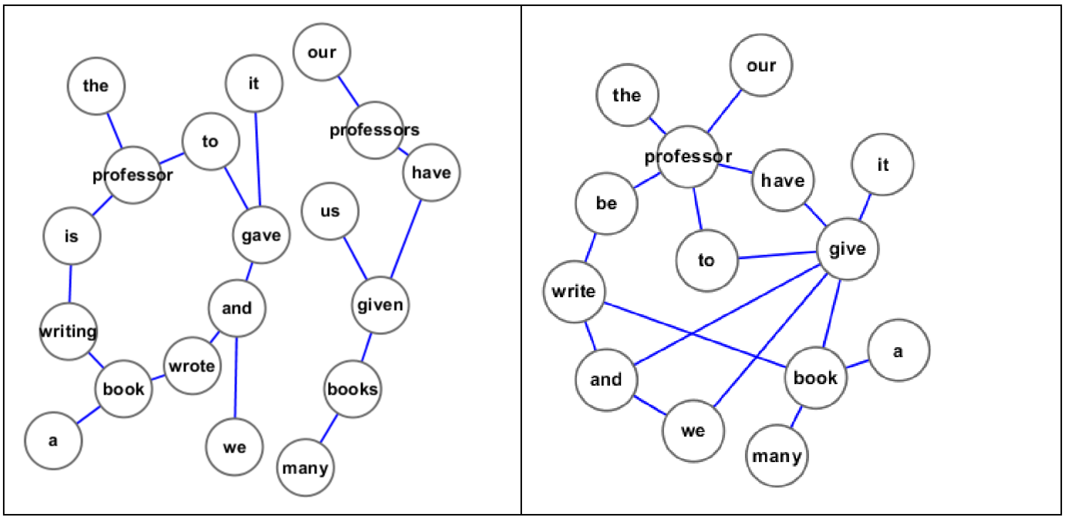
**Introduction. –** Languages are a complicate network structure [1]. Since traditional approaches of linguistic study can hardly research into the network properties of languages, linguists have to resort to new methods to study languages from a network perspective, which focuses on the overall picture of a language rather than the structural details. The complex-network approaches, based on empiricism and large-scale real corpora, facilitate the explorations into global properties of languages and advance our understanding of the complex human language structures. Besides, the application of theories of complex networks to linguistic study furthers the application of these theories to the fields of humanities and social sciences.

Scholars have conducted many studies regarding language and complex networks [2–10]. These studies involve many languages and adopt various principles in constructing language networks. These studies revealed that most language networks, though extracted from different languages and with diverse construction principles, all exhibit similar characteristics: scale free and small world. These researches are valuable for us to understand the universality of language networks. But so far, the complex-network approaches, which are overwhelming global-oriented, are rarely applied to the study of local and specific linguistic issues.

"Linguistic typology" is a discipline that concerns language classification. Traditional linguistic typology, which mainly depends on morphological features when classifying languages, is also called morphological typology. There is a good reason for morphological features to be taken as the parameters in language classification: the morphological variations can be mostly easily perceived. In the past, the typology studies seldom made use of large-scale real corpora since the technological means were rather limited. Recently, with the rapid development of information technology and large-scale real text processing technology, many studies concerning linguistic typology have been conducted on the basis of real texts [11–13].

Čech and Mačutek, after investigating lemma networks and word form networks of Czech, believe that the difference between them may reflect the typological features of a language [9]. Choudhury and Mukherjee found that the average degree of Hindi spelling network differs substantially from that of English, a discovery which may help linguists build a different linguistic typology theory [10]. Liu and Li constructed and researched the complex syntactic networks of 15 languages [14]. The results demonstrate that it is possible to classify human languages according to main parameters of complex networks with such

(a) E-mail: lhtzju@gmail.com




(a) (b)

Fig. 1: (Colour on-line) The word form network and the lemma network of three English sentences (a) is the word form network; (b) is the lemma network.

precision as can be achieved by means of modern word order typology.

Liu and Li have also found out that the study of language clustering on the basis of complex-network parameters is in fact the study of the degree of morphological variation from an overall perspective. Why do morphological variations —some micro phenomena— lead to the global differences between language networks? For the same language, what difference may exist between two syntactic networks which, respectively, take word form and lemma as nodes? If complex-network parameters reflect the degree of morphological variations, can the difference between word form network and lemma network better register the difference between languages? Will the observation and comparison of the two networks of one same language facilitate the discovery of emergent property of language networks? To address these questions, we built word form networks and lemma networks for 15 languages and extracted their network parameters. The second section introduces the methods and resources used in our research, the third section reports the main network parameters extracted from these networks and the fourth section presents discussions and conclusion, as well as suggestions for further researches.

**Methods and resources.** – Structurally, however big and complex a network may be, its elements are quite simple: nodes and edges. In different networks, nodes and edges represent different things. As far as syntactic networks are concerned, the nodes are word form or lemma and the edges are syntactic dependencies between them.

In our research, the dependency grammar is adopted to construct language networks. Dependency analysis is concerned with the binary relation between words and hence can be easily converted into a network representation [5].

Figure 1 presents syntactic networks of three exemplar English sentences: *This professor is writing a book; our professors have given us many books; we wrote a book and gave it to the professor.* These three sentences contain 23 word tokens; the word form network and the lemma network derived from these sentences, respectively, have 19 and 14 nodes.

In fig. 1, we can see *we* links to *and*, which links to *gave* and *wrote*. The reason is that we adopted, when annotating coordination, the annotation scheme used in Prague Dependency Treebank. There is evident difference between the above two networks, which is the ground on which we bring network study into linguistic typology.

Having constructed syntactic networks, we can research their major properties in terms of complex-network parameters. Average path length ($L$), cluster coefficients ($C$), average degree ($\langle k \rangle$), diameter ($D$), and degree distribution ($P(k)$) are the most frequently used parameters to determine the complexity of a network [15]. Considering the characteristics of a syntactic network, we also take the network centralization ($NC$) as a parameter [16]. Network centralization can help us find the central nodes of a syntactic network, which indirectly reflect the degree of morphological variations. Based upon these parameters, we can evaluate the properties of a network (*e.g.* whether it is a small-world or scale-free network).

For example, network in fig. 1(a) has the following parameters: $E$ (18), $N$ (19), $\langle k \rangle$ (1.895), $C$ (0), $L$ (2.713), $NC$ (0.069), $D$ (5), two connected components; and network in fig. 1(b) has the following parameters: $E$ (17), $N$ (14), $\langle k \rangle$ (2.429), $C$ (0.1), $L$ (2.462), $NC$ (0.321), $D$ (5), and one connected component. As to the distribution of nodes, there are, in fig. 1(a), 7 nodes whose degree is 1, 7 nodes whose degree is 2 and 5 nodes whose degree is 3; in fig. 1(b), there are 5 nodes whose degree is 1, 4 nodes whose degree is 2, 2 nodes whose degree is 3, 1 node whose degree is four, 1 node whose degree is five and 1 nodes whose degree is 6.

These data indicate that networks displayed in figs. 1(a) and (b) exhibit different properties of complex networks. Since the exemplar networks are extracted from only 3 sentences, 2 questions naturally arise: will these differences persist if we put more sentences (words) into networks? And if the answer is "yes", will these differences provide an answer to the question raised in the introduction?

To answer these 2 questions, we, with the available resources of treebanks, built dependency syntactic networks of the following 15 languages (ISO 639–2 language codes are in parentheses): Catalan (cat), Czech (cze), modern Greek (ell), ancient Greek (grc), Basque (eus), Hungarian (hun), Italian (ita), Portuguese (por), Spanish (spa), Turkish (tur), Latin (lat), Dutch (dut), French (fre), Slovenian (slv) and Russian (rus).

We used Network Analyzer [17], the network analysis plug-in of Cytoscape [18], to calculate the parameters of complex networks. Cytoscape is an open-source visualized bioinformatics software platform for molecular interaction network analysis. The results will be presented in the following section.

**The complexity of word form and lemma networks.** – Most of the treebanks used in our research come from the training set of CoNLL-X "Multi-language dependency syntactic analysis competition" [19,20]. All non-dependency treebanks have been converted by CoNLL-X's organizers into dependency ones, which are what we have used in this research. The details of these treebanks are available in the works listed in the reference [21–34]. We extracted from each treebank, as a sample, a continuous series of sentences that contains approximately the same amount of word tokens, and converted these samples into word form networks and lemma networks that can be analyzed with the complex-network analysis software.

We analyzed, with Network Analyzer, the two kinds of syntactic networks of all 15 languages. The results are shown in table 1.

Here, $E$ is the number of edges in the network; $N$ is the number of nodes; $\langle k \rangle$ is the average degree; $C$ is the cluster coefficients; $L$ is the average path length; $NC$ is the network centralization; $D$ is diameter; $\gamma$ is the power exponent of the degree distribution and $R^2$ is the determination coefficient of fitting the degree distribution to power law.

**Discussions of the comparisons between word form and lemma network.** – First of all, we compared the overall features of these networks, *i.e.*, the small-world and the scale-free features.

As can be seen in table 1, the fluctuation of the average path length of word form networks is not great, ranging from 2.958 to 3.938. The average path length of lemma networks fluctuates within an even narrower range: from 2.721 to 3.473. That is to say, the average distance between any two nodes will not exceed 3 nodes.

In a syntactic network, cluster coefficients reflect the possibility of a syntactic relationship between two words which are both syntactically related to another word. Our study reports that cluster coefficients of the word form networks range from 0.029 to 0.207 and those of the lemma networks fluctuate from 0.081 to 0.31. In comparison with random networks with the same nodes and the same average degree, we can see that the cluster coefficients of above two kinds of syntactic networks are much higher. Therefore, we may safely claim, in view of cluster coefficients and average path lengths in table 1, that the networks of the 15 languages under study are all small-world networks [35].

When the distribution of degrees in a network complies with the power law distribution ($P(k) \sim k^{-\gamma}$), the network is a scale-free one [36]. With the help of Network Analyzer, we carried out a power law fitting to the networks under study and obtained the power exponent and the determination coefficient $R^2$ of each language as shown in table 1.

The power exponents of word form networks range from 1.085 to 1.353 with only one language exhibiting a determination coefficient higher than 0.75. The power exponents of lemma networks fluctuate from 1.068 to 1.379 and the determination coefficients of eight languages exceed 0.75. The data demonstrates that, though the power exponent fluctuations of these two kinds of networks are rather similar, there is a better match between power law distribution and the degree distributions of lemma networks.

Our research also demonstrates that it is difficult to get convincing power law fitting results because the degree distribution of a real network characteristically has a long tail. Segmented fitting and accumulation of the degree distribution are commonly used to avoid the disturbance of a long tail. Researchers have proposed some new and more effective methods [37].

As shown in table 1, this parameter is enough to differentiate the languages under study and has the potentiality of becoming a parameter in language classification. According to the syntactic network researches conducted so far [2,5], when segmented fitting or accumulation of degree distribution are employed, the degree distributions of the networked explored in our research all follow a power law distribution. That is to say, all these networks are scale free.

After briefly viewing the overall features of these networks, we will observe some parameters that may be related to linguistic classification.

The degree of a node denotes the relations between a word and other words. Table 1 shows no relation between the average degree and the $NC$ of one language because $NC$ registers the differences among nodes in terms of their degrees, or the differences among the nodes regarding their ability to combine with other nodes,

Table 1: The main parameters of word form networks and lemma networks of 15 languages. For each language there are two rows of data, of which the upper one is the data of the word form network and the lower one is those of the lemma network.

| | $E$ | $N$ | $\langle k \rangle$ | $C$ | $L$ | $NC$ | $D$ | $\gamma$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| cat | 30944 | 8906 | 6.816 | 0.129 | 3.234 | 0.235 | 9 | 1.165 | 0.703 |
| | 27484 | 6089 | 8.725 | 0.236 | 2.875 | 0.366 | 8 | 1.117 | 0.738 |
| cze | 27447 | 10950 | 4.945 | 0.088 | 3.64 | 0.145 | 10 | 1.254 | 0.692 |
| | 23527 | 6070 | 7.534 | 0.157 | 3.24 | 0.2 | 8 | 1.247 | 0.764 |
| dut | 28873 | 9025 | 6.322 | 0.185 | 3.155 | 0.175 | 8 | 1.085 | 0.703 |
| | 26495 | 7457 | 6.966 | 0.233 | 3.016 | 0.201 | 8 | 1.068 | 0.685 |
| ell | 27942 | 9229 | 5.968 | 0.114 | 3.445 | 0.227 | 11 | 1.226 | 0.722 |
| | 22660 | 5182 | 8.485 | 0.237 | 2.923 | 0.386 | 8 | 1.195 | 0.757 |
| fre | 33169 | 8439 | 7.678 | 0.121 | 3.188 | 0.231 | 9 | 1.173 | 0.717 |
| | 27837 | 5939 | 8.971 | 0.195 | 2.913 | 0.38 | 8 | 1.154 | 0.747 |
| grc | 23798 | 8870 | 5.291 | 0.089 | 3.638 | 0.146 | 11 | 1.343 | 0.746 |
| | 17984 | 3682 | 9.389 | 0.187 | 3.105 | 0.231 | 7 | 1.214 | 0.812 |
| eus | 27895 | 10561 | 5.207 | 0.115 | 3.571 | 0.213 | 13 | 1.334 | 0.75 |
| | 21883 | 5124 | 8.233 | 0.242 | 3.054 | 0.295 | 9 | 1.198 | 0.795 |
| hun | 33146 | 13075 | 5.055 | 0.029 | 3.938 | 0.155 | 11 | 1.353 | 0.734 |
| | 28975 | 8607 | 6.672 | 0.081 | 3.473 | 0.199 | 9 | 1.379 | 0.769 |
| ita | 32329 | 9051 | 7.059 | 0.126 | 3.243 | 0.194 | 8 | 1.185 | 0.701 |
| | 27484 | 6089 | 8.725 | 0.236 | 2.875 | 0.366 | 8 | 1.117 | 0.738 |
| lat | 28945 | 11571 | 4.91 | 0.107 | 3.598 | 0.196 | 11 | 1.266 | 0.721 |
| | 23848 | 5305 | 8.644 | 0.191 | 3.114 | 0.265 | 8 | 1.239 | 0.804 |
| por | 29396 | 8855 | 6.444 | 0.207 | 3.123 | 0.312 | 8 | 1.125 | 0.685 |
| | 25509 | 6303 | 7.792 | 0.31 | 2.89 | 0.382 | 8 | 1.12 | 0.716 |
| rus | 42382 | 16543 | 5.088 | 0.091 | 3.55 | 0.176 | 12 | 1.203 | 0.696 |
| | 37309 | 8992 | 8.141 | 0.164 | 3.134 | 0.246 | 10 | 1.249 | 0.745 |
| slv | 19241 | 7128 | 5.309 | 0.125 | 3.473 | 0.171 | 9 | 1.164 | 0.700 |
| | 15832 | 4004 | 7.65 | 0.228 | 2.992 | 0.358 | 7 | 1.171 | 0.759 |
| spa | 25254 | 7939 | 6.209 | 0.181 | 3.146 | 0.271 | 9 | 1.108 | 0.688 |
| | 22180 | 5815 | 7.32 | 0.272 | 2.95 | 0.326 | 8 | 1.101 | 0.716 |
| tur | 26421 | 11969 | 4.25 | 0.205 | 2.958 | 0.514 | 10 | 1.161 | 0.616 |
| | 16296 | 3995 | 7.558 | 0.287 | 2.721 | 0.578 | 8 | 1.229 | 0.773 |

rather than the average ability of nodes to combine with other nodes. Syntactically, languages with high $NC$ have some nodes with extraordinarily high degrees. Researches of networks extracted from real texts reveal that these nodes are overwhelmingly function words or empty words. In other words, at least for word form networks, the more connections the function words have, the more synthetic this language is. Therefore, we may regard $NC$ as reflecting the degree of morphological variations and a seemingly useful parameter in language typology.

Theoretically, the average degree is related to the amounts of nodes and edges in a network, which motivated us to calculate the ratio between edges and nodes in each network. There is a strong correlation between the average degree and the ratio in both lemma networks and word form networks.

In linguistic typological researches based on real corpora, it is, when languages from the same family are devoid in samples, usually very difficult to ascertain those parameters truly independent of text size and annotation schemes because we can hardly judge whether the internal factors within languages or the non-linguistic factors should be responsible for a certain result.

In our samples, Italian, Portuguese, Catalan, Spanish and French belong to the Romance language subgroup whose ancestor is Latin. These languages are the reference languages from which we select parameters.

On the basis of preceding discussions, we take $\langle k \rangle$, $C$, $L$, $NC$, $D$, $\gamma$, and $R^2$ as variables. We used the clustering function provided by MiniTab to obtain the language cluster in terms of Euclidean minimum distance which is shown in fig. 2.

In the cluster of word form networks, the five romance languages fall into one group (79.65), though Dutch also belongs to it; the considerable resemblance (81.74) among Czech, Russian, Latin, modern Greek and ancient

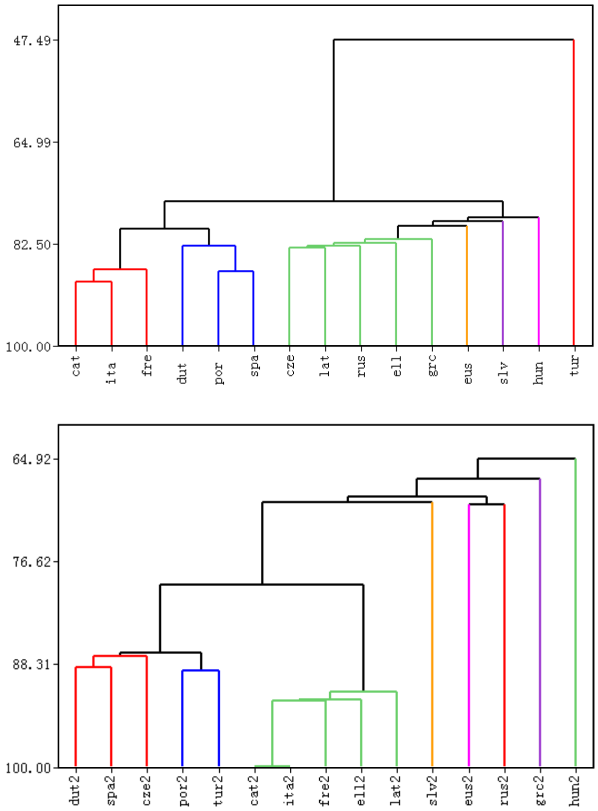



Fig. 2: (Colour on-line) Language clustering with 7 complex-network parameters. Top: the word form networks; bottom: the lemma networks.

Greek, which corresponds to their shared feature, *i.e*: rich inflections, betrays close relations among them.

A comparison between the graphs in fig. 2 and the graph in fig. 10 in [38] will show that the clustering results in this research, which is based on theories of complex network, are rather similar to the classifications reported in [38], which employs linguistic typological features in research. In other words, both of the methods are capable of distinguishing languages that are morphologically distinct.

Lemma networks, compared with word form networks, have the following features: edges and nodes are less; average degree and cluster coefficients are higher; average path lengths are shorter. These differences prove that the lemma network, compared with the word form network extracted from the same text, has a smaller size. In other words, the small-worldness of the lemma networks is more salient. At the same time, a higher determination coefficient implies that the distribution curve of node degrees of lemma networks has a better power law distribution fitting: of the 15 languages, determination coefficients of 8 are higher than 0.75 while, for word form networks, the determination coefficient of only one language (Basque) is higher than 0.75.

Through the comparison between lemma networks and word form networks, we can see that a better clustering result can be obtained from lemma networks than word

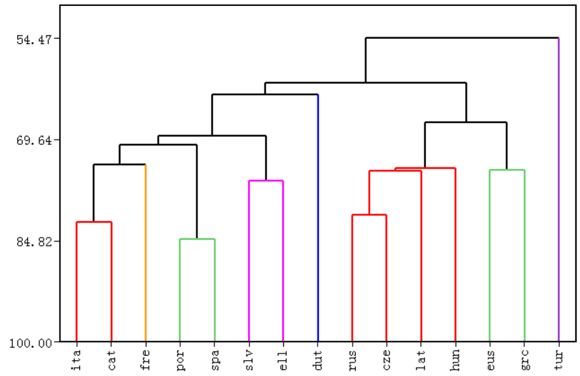



Fig. 3: (Colour on-line) Language clustering with seven parameters.

form networks when only five parameters (without $\langle k \rangle$ and $D$) are taken into account. However if seven parameters are taken into consideration, word form networks lead to better clustering result. This contrast may be due to different ranges of parameter variations of different languages. This issue cannot be sufficiently pursued here, but it is highly worthy of further researches. As a whole, lemma networks are more densely structured than word form networks. But different languages present different contraction degrees, a reflection of different morphological properties; although as Čech and Mačutek [9] have shown that such relation is not a straightforward reflection of different morphological properties. However, it seems reasonable to make a difference table of main parameters between word form networks and lemma networks of these 15 languages and investigate whether the differences may reflect the typological difference.

If the differences of these parameters between two networks reflect the degree of morphological variations in a language, it is reasonable to infer that languages of the same family should exhibit a similar degrees of morphological variation. Consequently, a clustering analysis on the basis of these differences may well gain a better result than the previous one. To test this hypothesis, we conducted more clustering experiments and found out that the best clustering result can be achieved with seven parameters. Figure 3 shows the result.

As shown in fig. 3, five romance languages are grouped in one cluster, though the resemblance level is only 70.5. This result agrees with the one obtained through approaches of modern language typology [11–13].

We also investigated whether language classification has any relation with the differences of average degrees and the differences of cluster coefficients between two kinds of networks. It is found that, though there is no high correlation, the languages can be more reasonably ordered in terms of the differences of average degree than cluster coefficients, a plausible evidence that, to a certain degree, supports [9].

According to the above discussions, it is obvious that word form networks can obtain better classification since lemma networks are devoid of any information of morphological variations. The clustering experiments also prove that the difference between lemma networks and word form networks is the best criterion in language classification.

**Conclusion.** – We explored into 15 word form networks as well as the corresponding lemma networks built on the basis of dependency syntactic treebanks, arriving at the finding that such network parameters as average degree, cluster coefficients, average path length, network centralization, diameter, power exponent of degree distribution and determination coefficient of fitting the degree distribution to power law, can, with similar accuracy as modern linguistic typological approaches can provide, classify the languages under study. Clustering experiments also show that word form networks can obtain better classification than lemma networks, which proves that language networks annotated with dependency schemas can, with the information of morphological variations embedded in them, classify languages from an overall perspective.

However, this new linguistic typology research method has its own defects that fall into two groups. The first group concerns the methods of complex-network research. The existing parameters of complex networks mostly focus on the global characteristics of a language and inevitably ignore the detailed difference of the language structure. Further works in this line should include adopting the social-network analysis technique, discovering new network parameters, and constructing weighted language networks. The second group concerns the corpora. Consistency should be secured in the corpora when language networks are constructed and the same dependency annotation scheme should be applied to samples of different styles of the same language or samples of different languages of the same style. On this basis, the commonness and individuality of these networks can be detected in comparative studies.

* * *

We thank the referees for insightful comments. This work was supported by the National Social Science Foundation of China (09BYY024) and partly supported by the Communication University of China as one of "211" Key Projects.

REFERENCES

[1] Hudson R., *Language Networks* (Oxford University Press, Oxford) 2007.
[2] Ferrer i, Cancho R., Solé R. V. and Köhler R., *Phys. Rev. E*, **69** (2004) 051915.
[3] Li J. and Zhou J., *Physica A*, **380** (2007) 629.
[4] Solé R. V. *et al.*, *Complexity*, **15** (2010) 20.
[5] Liu H., *Physica A*, **387** (2008) 3048.
[6] Liu H., *Chin. Sci. Bull.*, **54** (2009) 2781.
[7] Liu H. and Hu F., *EPL*, **83** (2008) 18002.
[8] Mehler A., *Corpus Linguistics.*, Vol. **1** (de Gruyter, Berlin, New York) 2008, pp. 328–382.
[9] Čech R. and Mačutek J., *Glottometrics*, **19** (2009) 85.
[10] Choudhury M. and Mukherjee A., *Dynamics On and Of Complex Networks* (Birkhäuser, Boston) 2009, pp. 145–166.
[11] Kelih E., *Glottometrics*, **20** (2010) 1.
[12] Bane M., *Proceedings of the 26th West Coast Conference on Formal Linguistics, 2008* (Cascadilla Proceedings Project, Somerville, Mass.) 2008, pp. 67–76.
[13] Popescu I.-I. and Altmann G., *J. Quant. Linguist.*, **15** (2008) 370.
[14] Liu H. and Li W., *Chin. Sci. Bull.*, **55** (2010) 3458.
[15] Albert R. and Barabási A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.
[16] Dong J. and Horvath S., *BMC Syst. Biol.*, **1** (2007) 24.
[17] Assenov Y. *et al.*, *Bioinformatics*, **24** (2008) 282.
[18] Shannon P. *et al.*, *Genome Res.*, **13** (2003) 2498.
[19] Buchholz S. and Marsi E., *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), New York, 2006*, pp. 149–164.
[20] Nivre J., *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007, Prague*, pp. 915–932.
[21] Aduriz I. *et al.*, *Second Workshop on Treebanks and Linguistic Theories, Vaxjo, Sweden, 2003.*
[22] Afonso S. *et al.*, *Proceedings of LREC-2002, Las Palmas, Canary Islands, Spain*, pp. 1698–1703.
[23] Atalay N. B., Oflazer K. and Say B., *Proceedings of LINC-2003, Budapest.*
[24] Bamman D. and Crane G., *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006), Prague*, pp. 67–78.
[25] Bamman D., Mambrini F. and Crane G., *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8), 2009*, pp. 5–15.
[26] Csendes D. *et al.*, *Proceedings of the 8th International Conference on Text, Speech and Dialogue, TSD 2005, LNAI 3658*, (Springer Verlag) 2005, pp. 123–131.
[27] Civit Torruella M. and Ma A. Marti Antonin, *Proceedings of TLT-2002, Sozopol, Bulgaria.*
[28] Montemagni S. *et al.*, *Treebanks*, edited by Abeillé A. (Kluwer, Dordrecht) 2003, pp. 189–210.
[29] Prokopidis P. *et al.*, *Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005), Barcelona*, pp. 149–160.
[30] Boguslavsky I. M. *et al.*, *Proceedings of the 18th Conference on Computational Linguistics, Saarbrücken, 2000*, Vol **2**, pp. 987–991.
[31] Abeillé A., Clément L. and Toussenel F., *Treebanks*, edited by Abeillé A. (Kluwer, Dordrecht) 2003.
[32] Dzeroski S. *et al.*, *Proceedings of LREC-2006, Genoa (Italy).*
[33] van der Beek L. *et al.*, *Computational Linguistics in the Netherlands* (Rodopi, Amsterdam), 2002.
[34] Hajic J., *Issues of Valency and Meaning* (Karolinum, Praha) 1998, pp. 106–132.
[35] Watts D. J. and Strogatz S. H., *Nature*, **393** (1998) 440.
[36] Barabási A.-L. and Albert R., *Science*, **286** (1999) 509.
[37] Clauset A., Shalizi C. R. and Newman M. E. J., *SIAM Rev.*, **51** (2009) 661.
[38] Liu H., *Lingua*, **120** (2010) 1567.