Article
Chinese Science Bulletin
Applied Physics
April 2013 Vol.58 No.10: 1139–1144
doi: 10.1007/s11434-013-5711-8

Language clustering with word co-occurrence networks based on parallel texts

LIU HaiTao* & CONG Jin

School of International Studies, Zhejiang University, Hangzhou 310058, China

Received September 13, 2012; accepted November 15, 2012

This study investigates the feasibility of applying complex networks to fine-grained language classification and of employing word co-occurrence networks based on parallel texts as a substitute for syntactic dependency networks in complex-network-based language classification. 14 word co-occurrence networks were constructed based on parallel texts of 12 Slavic languages and 2 non-Slavic languages, respectively. With appropriate combinations of major parameters of these networks, cluster analysis was able to distinguish the Slavic languages from the non-Slavic and correctly group the Slavic languages into their respective sub-branches. Moreover, the clustering could also capture the genetic relationships of some of these Slavic languages within their sub-branches. The results have shown that word co-occurrence networks based on parallel texts are applicable to fine-grained language classification and they constitute a more convenient substitute for syntactic dependency networks in complex-network-based language classification.

word co-occurrence network, Slavic languages, parallel texts, language classification, cluster analysis

Citation: Liu H T, Cong J. Language clustering with word co-occurrence networks based on parallel texts. Chin Sci Bull, 2013, 58: 1139–1144, doi: 10.1007/s11434-013-5711-8

Complex networks are ubiquitous and pervade in almost all facets of the natural world and human activity [1]. In recent years, complex networks have started to be applied to both theoretical and practical research concerning human language [2]. The conception of language as a system is one of the central assumptions of modern linguistics [3]. If language is a system, its organization patterns at the system level cannot be adequately represented and characterized by traditional methods in linguistics, which in turn focus on the fine detail of language structure. Complex networks, which enable a holistic view of various systems, can fill the methodological gap in linguistics. Different aspects and levels of human language can be modeled and characterized as linguistic networks [4–6], with relevant linguistic units as vertices (or nodes) and their relationships of a particular type as edges (or links).

Quantitative analysis of linguistic networks can be translated into potential methods in different fields in linguistics. Language classification is a representative example. Studies [7–9] have shown that we can classify languages through cluster analysis of their syntactic dependency networks (with different word forms as vertices and the syntactic dependency relations between them as edges) according to their major complex network parameters. The results of classification can generally capture the genetic relationships of the languages as found in the language families. This complex-network-based language classification falls under the heading of typological classification, which focuses on structural features of languages [10]. Liu and Xu's findings [8] suggested that the major complex network parameters of a syntactic dependency network are indicators mainly of the morphological and syntactic properties of the corresponding language at the system level. Therefore, complex-network-based language classification is an important contribution to holistic typology [11]. The feasibility of the complex-network-based language classification [7–9] indicates that the major parameters of complex networks can capture the diversity of networks in the real world, in addition to re-

*Corresponding author (email: lhtzju@gmail.com)

vealing their commonality (e.g. the non-trivial statistical patterns universally found in various networks, including small-world and scale-free properties). Meanwhile, the use of complex networks in language classification also expands the application of complex networks and broadens the horizon of complex networks research.

It is noteworthy that the previous studies of complex-network-based language classification [7–9] were usually satisfied with roughly classifying the languages into their respective branches (e.g. Romance, Germanic and Slavic) without considering the subdivision of these branches. That is to say, there has been no study that is devoted to rather fine-grained language classification (e.g. how languages of the same language branch are subdivided into different sub-branches) from the complex-network-based approach. If the application of complex networks can be proved to be able to yield fine-grained language classification, linguistic typology can benefit more from this complex-network-based approach and the application of complex networks can also be expanded to more specific fields in humanities and social sciences. Methodologically speaking, there are two major problems of the complex-network-based language classification in these previous studies. On the one hand, the syntactic dependency networks in these studies were based on language data which were not necessarily consistent in semantic content and genre. The basic assumption of language classification based on syntactic dependency networks is that the topological similarities and differences of these networks (manifested by their complex network parameters) reflect the similarities and differences of the corresponding languages [9]. However, the inconsistency in semantic content and genre of the language data selected, which is independent of the similarities and differences of the languages, may also contribute to the topological similarities and differences of the corresponding syntactic dependency networks and thus may affect the results of language classification. A more desirable type of language data for complex-network-based language classification is parallel texts (i.e. a collection of texts with the same semantic content but in different languages, e.g. a novel plus its translations in different languages), which are consistent in both semantic content and genre. On the other hand, the construction of syntactic dependency networks requires considerable manpower and material resources. Syntactic dependency networks are derived from syntactic dependency treebanks. The latter are based on annotation of the raw language data with syntactic dependency. Although some methods of automatic annotation are available, the annotation has to be conducted manually in a word-by-word and sentence-by-sentence manner in order to achieve satisfactory accuracy for linguistic research. Therefore, even though syntactic dependency networks can constitute an effective method for language classification, it is hard to apply this method to classification of a rather large number of languages considering the difficulty of the construction of syntactic dependency networks. In addition, the difference in approaches to syntactic dependency annotation may also affect the topological properties of the syntactic dependency networks and thus the results of language classification. Therefore, we need to find a more available type of linguistic network as an alternative to syntactic dependency networks. Of all other types of linguistic network, word co-occurrence networks [12] can be a candidate to fill this role (see Section 1 for detailed introduction). In view of the above two problems, we can consider employing word co-occurrence networks based on parallel texts as a potential substitute for syntactic dependency networks in complex-network-based language classification.

This study investigates the feasibility of applying complex networks to fine-grained language classification and of employing word co-occurrence networks based on parallel texts as a substitute for syntactic dependency networks in complex-network-based language classification. We constructed 14 word co-occurrence networks based on parallel texts of 12 Slavic languages and 2 non-Slavic languages, respectively, and conducted cluster analysis to these networks according to different combinations of their major complex network parameters. The effect of classification was evaluated through comparison of the results of clustering against the genetic relationships of these languages (especially the 12 Slavic languages) as found in the language families.

1 Methods and materials

A word co-occurrence network, which is derived from a body of authentic language data, can be defined and thus constructed in more than one way. In this study we define "co-occurrence" as the adjacency of two word forms in sentence formation. For instance, in "John kicked the ball" there are three pairs of adjacent word forms, namely "John kicked", "kicked the" and "the ball". A word co-occurrence network thus can be represented by an undirected graph $G = (V, E)$. V is the set of vertices representing all the different word forms in the language data. E, on the other hand, is the set of edges representing all different adjacency relations of the word forms in sentence formation. Therefore, two vertices $u, v \in V$ are joined by an edge $e \in E$ if the two corresponding word forms are adjacent within at least one sentence. According to this definition, we can extract all the different word-form bigrams in sentence formation from the authentic language data and convert this set of bigrams into the word co-occurrence network. A word co-occurrence network can be constructed automatically. A major advantage of using word co-occurrence networks lies in their unambiguity, for a co-occurrence relation can be unequivocally defined and extracted from the language data in a theory-neutral manner. Figure 1 displays a word co-occurrence network constructed according to the above definition (the

language data were excerpted from Chapter 1 of Steven Pinker's book *The Language Instinct: The New Science of Language and Mind*). Unless otherwise specified, a word co-occurrence network refers to one constructed according to the above definition for the rest of this paper.

A word co-occurrence network and a syntactic dependency network, suppose they are derived from the same body of authentic language data, differ only in the type of edges. The edges of the former represent the adjacency relations of the word forms in sentence formation, whereas those of the latter the syntactic dependency relations in sentence formation. Statistics from a number of different languages [13] have shown a high probability (usually over 50%) for a syntactic dependency relation to be between two adjacent word forms. This means that a word co-occurrence network and its syntactic dependency network counterpart derived from the same body of authentic language data tend to be highly similar in terms of network topology, for there is a significant overlap between the edges of the two types of network. For instance, the central vertices of the word co-occurrence network in Figure 1 tend to be function words, which is consistent with the case of a syntactic dependency network [14,15]. Therefore, word co-occurrence networks can constitute a potential substitute for syntactic dependency networks in studies of linguistic networks. The complex network parameters of a word co-occurrence network can be adopted as a convenient approximation for those of its syntactic dependency network counterpart as indicators of the morphological and syntactic properties of a language at the system level.

The parallel texts on which the word co-occurrence networks in this study were based are of the following 14 languages: Russian, Belarusian, Ukrainian, Czech, Slovak, Polish, Upper-Sorbian, Serbian, Croatian, Slovenian, Bulgarian, Macedonian, English and Chinese. Of these 14 languages, 12 are Slavic languages, which fall into three subbranches, namely Eastern (Russian, Belarusian and Ukrainian), Western (Czech, Slovak, Polish and Upper-Sorbian) and Southern (Serbian, Croatian, Slovenian, Bulgarian and Macedonian) [16]. These parallel texts are the novel "How the steel was tempered" (Kak zakaljalas' stal') in the Russian original (written by N.A. Ostrovskij in the years 1932–1934) and its translations into the other 13 languages. The parallel texts of the 12 Slavic languages are from the Slavic parallel corpus constructed by Emmerich Kelih (for detailed introduction of the corpus see [17]), whereas the English and Chinese texts are what we obtained from the translations of the novel in these two languages. As most of these

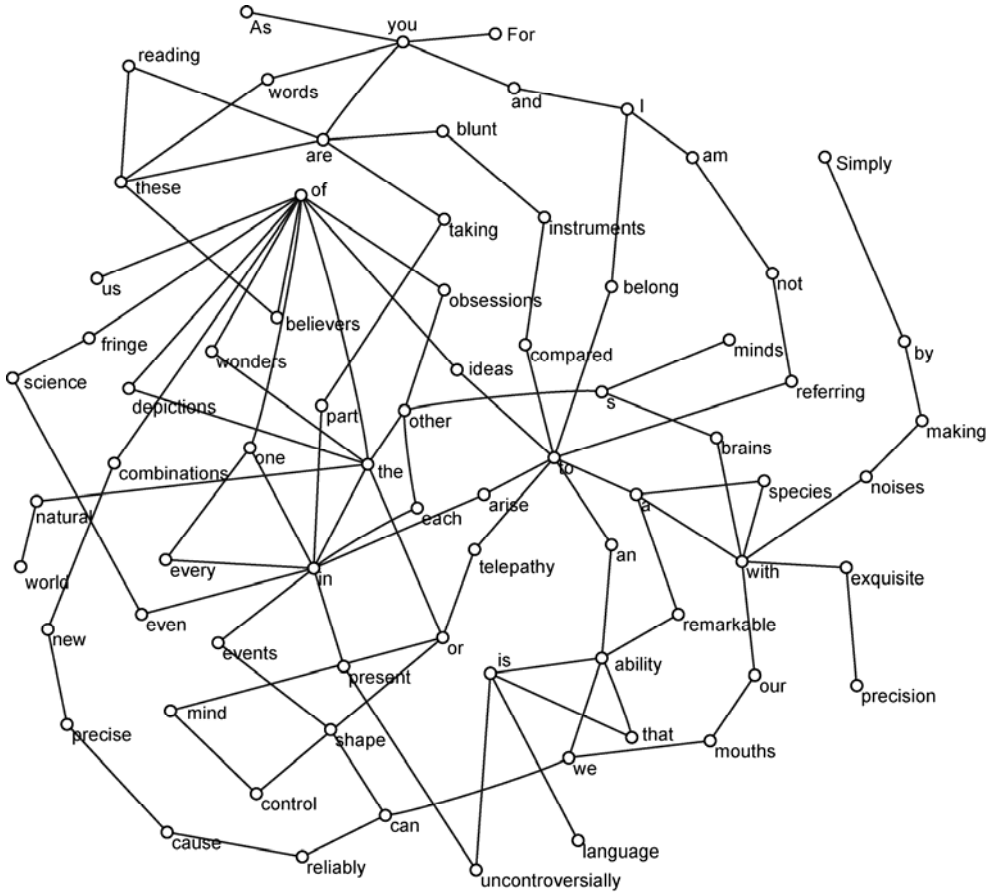

Figure 1 A word co-occurrence network of English.

languages belong to the Slavic branch, which in turn can be subdivided into different sub-branches, it is possible to examine the effect of applying word co-occurrence networks based on parallel texts to fine-grained language classification.

We employed NetworkAnalyzer [18], one of the plugins of Cytoscape, a platform for complex network analysis, in the calculation of the 10 complex network parameters of the 14 word co-occurrence networks. These complex network parameters are average degree (<*k*>), average path length (*L*), clustering coefficient (*C*), network centralization (*NC*), diameter (*D*), network heterogeneity (*NH*), exponent of the power law of best fit to *P*(*k*) (i.e. degree distribution) (γ_1), coefficient of determination for the power law of best fit to *P*(*k*) (R_1^2), exponent of the power law of best fit to $\overline{k}_{nn}(k)$ (i.e. the distribution of average nearest neighbors degree) (γ_2), and coefficient of determination for the power law of best fit to $\overline{k}_{nn}(k)$ (R_2^2) (for detailed account of the parameters and their applications see [9,19]).

These above parameters are sufficient to provide us a general picture of a complex network's topological properties, for instance, whether it is a small-world or scale-free network. The use of cluster analysis in language classification can be traced back at least to the work of Altmann and Lehfeldt [20]. Cluster analysis was conducted to the 14 networks according to different combinations of the parameters calculated. These parameters were standardized before they were inputted into the cluster analysis. Ward method and Manhattan distance were adopted for the cluster analysis. Based on the experience of previous research concerning complex-network-based language classification [7–9], we selected the combination of <*k*>, *L*, *C* and *NC* as a base set. Other combinations were formed by adding other parameters to the base set. Altogether 64 combinations were checked in the cluster analysis.

2 Results and discussion

With the method introduced in Section 1, we obtained the major parameters of the 14 word co-occurrence networks, which in turn are displayed in Table 1.

The effect of classification was evaluated through comparison of the results of clustering against the genetic relationships of these languages as found in the language families. As the 14 languages are mostly Slavic languages, we focused on how well the results of clustering captured the genetic relationships of the 12 Slavic languages. The basic criterion for evaluating the effect of classification is that the 12 Slavic languages must be clustered together before they were clustered with the 2 non-Slavic languages. In other words, the results of clustering must be able to distinguish the Slavic languages from the non-Slavic. This criterion satisfied, we checked whether the 12 Slavic languages were clustered correctly into their respective sub-branches.

Of the 64 combinations of complex network parameters checked, 15 yielded a result which could distinguish the Slavic languages from the non-Slavic and correctly group the 12 Slavic languages into their respective sub-branches. Illustrated in Figure 2 is one of these results, which was yielded with the base set plus *D*, R_1^2, γ_2 and R_2^2. Figure 2 well demonstrates the subdivision of the Slavic languages, for the 12 Slavic languages were correctly grouped into their respective sub-branches. In addition, the clustering could also capture the genetic relationships of some of these Slavic languages within their sub-branches. For instance, although Serbian and Croatian adopt different writing systems, it is commonly accepted that they are the same language [16]. As demonstrated by Figure 2, Serbian and Croatian were clustered together at a distance of 1.70 in their sub-branch. The close genetic similarity between Bulgarian and Macedonian was also captured (at a distance of 3.57). This result of the classification of the Slavic languages is

Table 1 Major parameters of the word co-occurrence networks of 14 languages

	$\langle k \rangle$	*L*	*C*	*NC*	*D*	*NH*	γ_1	R_1^2	γ_2	R_2^2
Belarusian	4.819	3.797	0.100	0.114	17	5.833	1.232	0.742	0.451	0.794
Bulgarian	5.690	3.354	0.186	0.144	11	6.767	1.159	0.711	0.525	0.855
Chinese	8.684	2.944	0.283	0.354	9	6.113	1.180	0.755	0.534	0.930
Croatian	5.353	3.479	0.151	0.127	13	6.574	1.212	0.712	0.505	0.847
Czech	4.945	3.627	0.119	0.157	13	6.696	1.257	0.75	0.500	0.873
English	9.043	2.964	0.299	0.297	10	5.499	1.157	0.743	0.533	0.883
Macedonian	6.206	3.225	0.220	0.170	10	6.698	1.138	0.724	0.546	0.841
Polish	4.983	3.628	0.118	0.112	14	6.351	1.229	0.720	0.475	0.824
Russian	4.504	3.891	0.091	0.109	17	5.972	1.268	0.748	0.444	0.757
Serbian	5.348	3.485	0.147	0.126	15	6.543	1.213	0.707	0.515	0.832
Slovak	5.166	3.592	0.128	0.137	14	6.255	1.235	0.747	0.477	0.836
Slovenian	5.367	3.406	0.164	0.192	13	7.400	1.192	0.738	0.565	0.787
Ukrainian	4.865	3.814	0.096	0.076	16	5.433	1.254	0.764	0.424	0.737
Upper-Sorbian	5.347	3.550	0.131	0.161	14	6.359	1.239	0.741	0.466	0.822

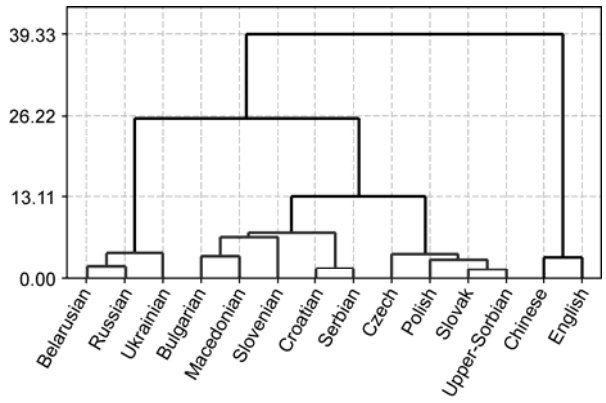

Figure 2 Clustering of the 14 word co-occurrence networks with 8 complex network parameters.

more straightforward than that achieved by examining the type-token relationship in Slavic languages based on the same Slavic parallel corpus in Kelih's study [17], which in turn only yielded a rank of the 12 Slavic languages reflecting their genetic closeness without suggesting how the languages should be classified. This result is also generally comparable with those achieved by other methods including lexicostatistics [21].

The cluster analysis also involved English and Chinese, which are non-Slavic. As illustrated in Figure 2, English and Chinese as one cluster and the 12 Slavic languages as the other were clustered at a distance of 39.33, whereas English and Chinese were clustered at a distance of 3.34. This result reflects not only the difference between English and Chinese as non-Slavic languages and the 12 Slavic languages, but also the close similarity between English and Chinese as also found in previous research based on authentic language data [7,22].

The method adopted in this study is highly automatic with little human aid. For instance, it is not necessary to bother about the writing systems of the languages and it has also been proved that the difference in writing system does not contribute to the clustering of the languages in question. Of the 12 Slavic languages, Russian, Belarusian, Ukrainian, Serbian, Bulgarian and Macedonian adopt Cyrillic alphabet, while the other 6 adopt Latin alphabet. However, their difference in writing system does not contribute to their grouping as indicated by the result illustrated in Figure 2. This also inspires our consideration of the relationship between language and writing system. For example, the Chinese language seems to be totally different from English, judging from its peculiar writing system. However, as indicated by the result of this study together with those of [7,22], the difference between the two languages turns out to be much smaller than expected. It is also noteworthy that the effect of classification for Slavic languages in this study is significantly better than that based on such word-order parameters as dependency direction in Liu's study [22]. This is because the method adopted in this study relies on global characterization of language as a system, instead of a set of local structural details, which can hardly capture the wholeness of the language system. This also indicates that word order may not be the most appropriate basis for the classification of languages with richer inflectional morphology as in the case of Slavic languages [23]. In addition, as the method in this study approaches language classification totally in quantitative terms, the similarities and differences of the languages which it can capture are continuous rather than discrete.

3 Conclusions

This study investigates the feasibility of applying complex networks to fine-grained language classification and of employing word co-occurrence networks based on parallel texts as a substitute for syntactic dependency networks in complex-network-based language classification. We constructed 14 word co-occurrence networks based on parallel texts of 12 Slavic languages and 2 non-Slavic languages, respectively, and conducted cluster analysis to them according to different combinations of their major complex network parameters. With appropriate combinations of these parameters, cluster analysis was able to distinguish the Slavic languages from the non-Slavic and correctly group the Slavic languages into their respective sub-branches. Moreover, the clustering could also capture the genetic relationships of some of these Slavic languages within their sub-branches. Therefore, a conclusion can be drawn that word co-occurrence networks based on parallel texts are applicable to fine-grained language classification and they constitute a more convenient substitute for syntactic dependency networks in complex-network-based language classification. The methodology adopted in this study can also help to establish a holistic and quantitative approach to linguistic typology which can capture the continuous similarities and differences of languages. This study further confirms the feasibility of applying the major complex network parameters to probing into the diversity of real-world networks. More importantly, as parallel-text-based word co-occurrence networks have been proved to be able to handle fine-grained language classification, the application of complex networks can be expanded further to more specific fields in humanities and social sciences.

We thank the anonymous reviewers for valuable suggestions and insightful comments and Emmerich Kelih for providing the Slavic parallel texts used in this study. This work was supported by the National Social Science Foundation of China (*09BYY024 and 11&ZD188*).

1 Costa L D F, Oliveira O N, Travieso G, et al. Analyzing and modeling real-world phenomena with complex networks: A survey of applications. Adv Phys, 2011, 60: 329–412
2 Choudhury M, Mukherjee A. The structure and dynamics of linguistic networks. In: Dynamics on and of Complex Networks, Modeling

and Simulation in Science, Engineering and Technology. Boston: Birkhaeuser, 2009. 145–166
3 Kretzschmar W A. The Linguistics of Speech. New York: Cambridge University Press, 2009
4 Steyvers M, Tenenbaum J B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognit Sci, 2005, 29: 41–78
5 Ferrer i Cancho R, Solé R V, Köhler R. Patterns in syntactic dependency networks. Phys Rev E, 2004, 69: 051915
6 Liu H T. Statistical properties of Chinese semantic networks. Chin Sci Bull, 2009, 54: 2781–2785
7 Liu H T, Li W W. Language clusters based on linguistic complex networks. Chin Sci Bull, 2010, 55: 3458–3465
8 Liu H T, Xu C S. Can syntactic networks indicate morphological complexity of a language? Europhys Lett, 2011, 93: 28005
9 Abramov O, Mehler A. Automatic language classification by means of syntactic dependency networks. J Quant Ling, 2011, 18: 291–336
10 Ruhlen M. A Guide to the World's Languages 1: Classification. Stanford: Stanford University Press, 1991
11 Shibatani M, Bynon T. Approaches to language typology: A conspectus. In: Approaches to language typology. New York: Oxford University Press, 1995. 1–26
12 Ferrer i Cancho R, Solé R V. The small world of human language. Proc R Soc Lond B, 2001, 268: 2261–2265
13 Liu H T. Dependency distance as a metric of language comprehension difficulty. J Cognit Sci, 2008, 9: 159–191
14 Solé R V, Corominas-Murtra B, Valverde S, et al. Language networks: Their structure, function and evolution. Complexity, 2010, 15: 20–26
15 Chen X Y, Liu H T. Central nodes of the Chinese syntactic networks (in Chinese). Chin Sci Bull (Chin Ver), 2011, 56: 735–740
16 Katzner K. The Languages of the World (New Edition). London and New York: Routledge, 1995
17 Kelih E. The type-token relationship in Slavic parallel texts. Glottometrics, 2010, 20: 1–11
18 Assenov Y, Ramirez F, Schelhorn S E, et al. Computing topological parameters of biological networks. Bioinformatics, 2008, 24: 282–284
19 Costa L D F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements. Adv Phys, 2007, 56: 167–242
20 Altmann G, Lehfeldt W. Allgemeine Sprachtypologie. Munich: Fink, 1973
21 Novotná P, Blažek V. Glottochronolgy and its application to the Balto-Slavic languages. Baltistica, 2007, XLII: 185–210
22 Liu H T. Dependency direction as a means of word-order typology: A method based on dependency treebanks. Lingua, 2010, 120: 1567–1578
23 Comrie B, Corbett G G. Introduction. In: The Slavonic Languages. London: Routledge, 2002. 1–19