Available online at www.sciencedirect.com
ScienceDirect
Physica A 387 (2008) 3048–3058
PHYSICA A
www.elsevier.com/locate/physa

The complexity of Chinese syntactic dependency networks

Haitao Liu*

Institute of Applied Linguistics, Communication University of China, China

Received 6 September 2007; received in revised form 15 December 2007
Available online 18 January 2008

Abstract

This paper proposes how to build a syntactic network based on syntactic theory and presents some statistical properties of Chinese syntactic dependency networks based on two Chinese treebanks with different genres. The results show that the two syntactic networks are small-world networks, and their degree distributions obey a power law. The finding, that the two syntactic networks have the same diameter and different average degrees, path lengths, clustering coefficients and power exponents, can be seen as an indicator that complexity theory can work as a means of stylistic study. The paper links the degree of a vertex with a valency of a word, the small world with the minimized average distance of a language, that reinforces the explanations of the findings from linguistics.

PACS: 89.75.Da; 89.75.Fb

Keywords: Chinese syntactic dependency network; Complexity; Average path length; Clustering coefficient; Degree distribution; Genre

1. Introduction

Viewing a language as a network or a system with interconnected elements is not new to contemporary linguistics [1,2]. However, if we investigate the achievements of linguistics from the perspective of network science, it is still unrefined, but the structure of language, which can be observed from a network's view, makes it possible to investigate a language with modern network science.

Recently, network science has had essential developments [3–5]. Many scholars investigate different language networks [6,7]. The researchers often build language networks based on the following principles: (1) relations between the root word and its synonyms in a thesaurus; (2) word-association relation in a Wordnet-like lexicon; (3) co-occurrence in a sentence; (4) syntactic relations in a treebank. For Chinese, Refs. [8,9] investigates some properties of phrase networks, which are built based on the same Hanzi (Chinese character) among the phrases; Ref. [10] presents some findings based on Chinese character networks.

The above-mentioned studies show that language networks are just like most other real-world networks, which are small-world and scale-free. These studies are useful for understanding the generality of language structure, but it is a pity that some methods lack credible linguistics supports, which makes the results difficult to be explained

* Tel.: +86 1065783456; fax: +86 1065783456.
E-mail address: lhtcuc@gmail.com.

0378-4371/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2008.01.069

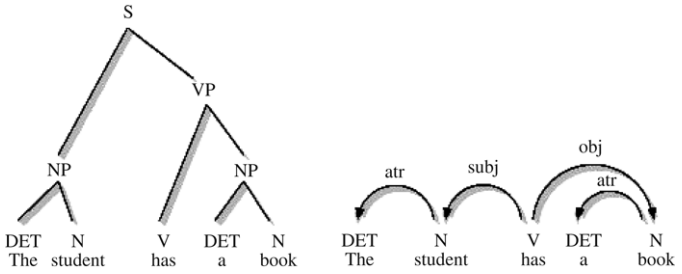

Fig. 1. Comparison of constituency analysis to dependency analysis.

from linguistics. Without good explanations, the studies would be not very helpful to explore the dynamics of the network and linguistics. Then, it seems necessary to build a language network based on linguistic theories, because for linguists, networks are means, not the goal.

The paper analyzes two Chinese syntactic dependency networks, which are built from two dependency treebanks with different genres. Compared with other studies, our networks have clearer linguistic motivation and methods to build the language network based on syntactic principles.

In the paper, we try to compare the results of network analysis of two syntactic networks with previous studies on complex networks of linguistic units. Two networks with different genres are used with the target of observing if the properties of complex networks can be used as a means to distinguish the genres in particular, or network analysis as one of the means of linguistic research, in general.

Section 2 introduces the resources building Chinese syntactic dependency networks. Section 3 discusses the results of network analysis. Section 4 is concluding remarks.

2. Resource and method

Networks, in particularly real networks, are mostly complex networks. However, complex does not mean that the elements of networks are also complex. All networks consist of edges and vertices, which represent different things in real networks. In a language network, a vertex can be different linguistic units, for example, parts of Chinese characters, Chinese characters, and words, while the edges describes the relation between these units.

The focus of the paper is syntactic networks, in which a vertex is often a word (type), and the edge is the relation between two words. It seems that researchers generally use the word (type) as vertex of a syntactic network without doubts, but like to use various means to build links between two words. Syntax is a science how to analyze and organize a sentence. In this way, while we construct a syntactic network, we should follow a syntactic theory. Otherwise, the built network will not be sufficiently convincing, at least if the study is linguistically oriented. For instance, a statistic based on syntactic corpora with 20 languages in Ref. [11] shows that only about 50% of adjacent words are syntactically relevant. Therefore, it is not very sound to build a syntactic network of human languages only based on adjacent co-occurrence [12].

Constituency and dependency analyses are two principal methods for getting the syntactic representation of a sentence. In constituency (or phrase structure) analysis, structures reveal how linguistic units form a greater unit. It is a part-whole analysis. Contrastingly, in dependency analysis, structures consist of binary asymmetrical relations between words in a sentence. The difference between two approaches are shown in Fig. 1.

It is obvious that dependency analysis is more network-friendly, because a dependency structure is based on such elements which are also found in a network.

The following properties, which are generally accepted by linguists, are considered the core features of a syntactic dependency relation [2,13]:

1. It is a binary relation between two linguistic units.
2. It is usually asymmetrical, with one of the two units acting as the governor and the other as dependent.
3. It is labeled, so the dependency relations should be distinguished and explicitly labeled in the arc linking the two units.

They form a dependency relation as shown in Fig. 2.

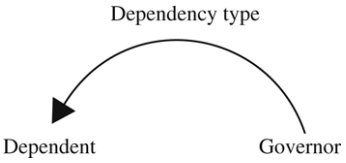

Fig. 2. Three elements of a dependency.

Table 1
Annotation of two English sentences in the treebank

Order number of sentence	Dependent			Governor			Dependency type
	Order number	Word	POS	Order number	Word	POS	
S1	1	the	det	2	student	n	atr
S1	2	student	n	3	has	v	subj
S1	3	has	v				
S1	4	a	det	5	book	n	atr
S1	5	book	n	3	has	v	obj
S2	1	he	pr	2	reads	v	subj
S2	2	reads	v				
S2	3	an	det	5	book	n	atr
S2	4	interesting	adj	5	book	n	atr
S2	5	book	n	2	reads	v	obj

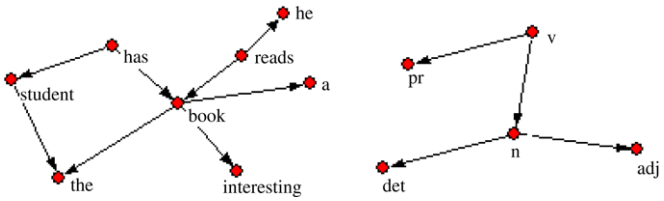

Fig. 3. An example of a syntactic dependency network.

Fig. 2 show a dependency between *Dependent* and *Governor*, whose label is *Dependency type*. The directed arc from *Governor* to *Dependent* demonstrates the asymmetrical relation between the two units. Dependency analysis can be seen as a set of all dependencies found in a sentence. From the network's view, a dependency constitutes a minimum unit which constitutes a network, i.e. Dependent and Governor are vertices, and the dependency link between them is edge.

Treebanks are a corpus with syntactic annotation. They are often used as a tool and a resource for training and evaluating a syntactic parser in computational linguistics [14]. Table 1 shows the analysis of two English sentences in terms of the dependency syntax with each word token distinguished by a number showing the linear order of the word in the sentence.[1]

In Table 1, each row is a dependency relation including three elements: dependent, governor and dependency type. All rows in a sentence consist of the dependency structure. A sentence with n words includes n-1 dependencies. It is possible to convert a dependency analysis into a network graph, while a dependency treebank is converted into a graph, which forms a syntactic dependency network. A dependency relation in treebank is converted into a link in a network. Fig. 3 shows a syntactic dependency network including two English sentences in Table 1.

The left of Fig. 3 is a syntactic dependency network based on the columns "Word" of dependent and governor in Table 1. The right is a syntactic dependency network based on the columns "POS", which consists of word classes in a treebank. The right one can be used as resource to explore the network's properties of word classes in a language. In this paper, we only investigate the left type. Fig. 3 also shows that the vertices of a syntactic dependency network are the word types, which make and link the separated network-fragments based on a sentence into a greater network of

[1] In Table 1, *det* is a determiner, *v* is a verb, *n* is a noun, *adj* is an adjective, *pr* is a pronoun. *subj* is subject, *atr* is attribute, *obj* is object.

the treebank. For instance, in the treebank, the word type *Book* appears twice as a dependent and thrice as a governor, therefore, there are five edges linking the vertex *Book* to other vertices in the network.

Compared with the original dependency grammar, this is a way which can be used for exploring the strength of the links between nodes in the grammar. For example, if ten tokens of the word type '*book*' are linked directly to tokens of 'read', this will presumably appear as ten separate links in the network, registering the frequency of this pairing. In a syntactic network, all the links are dependencies that are permitted by the grammar. However, in order to allow comparison with the previous analyses, in this paper we have ignored such strength consideration and converted multiple identical links into a single link. Thus, although each token contributes a separate link to the total network, the degrees that are calculated later are only a measure of connectedness.

One of two treebanks used to construct the syntactic networks is built on the news (xinwen lianbo, hereinafter "xwlb") of China Central Television, a genre which is intended to be spoken but whose style is similar to the written language. The treebank xwlb includes 16654 word tokens, so the mean sentence length is 24 words. Another used treebank is based on a conversational program (shihua shishuo, hereinafter "shss") of China Central Television, which has 19060 word tokens and the mean sentence length 21.

Following the procedures above, the generated networks are directed networks. For comparing with the previous studies, we have to convert them into undirected networks.

We use the software Pajek [15] and Minitab [16] to analyze two syntactic networks. The results and discussion are in the next section.

3. Analysis of Chinese syntactic dependency networks

For evaluating the complexity of a network, the most often investigated network indicators are its average path length, clustering coefficients and degree distribution.

Average path length $\langle d\rangle$ is defined as the average shortest distance between any pair of vertices in a network:

$$\langle d\rangle = \frac{1}{\frac{1}{2}N(N+1)}\sum_{i\geq j} d_{ij}. \tag{1}$$

Here, N is the number of vertices in the network; d_{ij} is distance between the vertices i and j, which can be defined as the number of edges in a shortest path linking the two vertices.

Diameter D is defined as the longest shortest path in a network. For instance, in the xwlb syntactic network, the longest shortest path is found from the vertex 821 to vertex 3032, because there are 10 edges in this path, thus, the diameter of the network is 10.

In a syntactic network, the number of links of a given word type is called its degree k. $\langle k\rangle$ is average degree of a network. Degree distributions are defined as the frequency $P(k)$ of having a word type with k links.

Clustering coefficient is the probability that two vertices (e.g. word types) that are neighbors of a given vertex are neighbors of each other. Let k_i denotes degree of vertex i, E_i denotes the number of edges among the vertices in the nearest neighborhood of vertex i. Then, the clustering coefficient C_i of the vertex i is defined as [4]:

$$C_i = \frac{2E_i}{k_i(k_i-1)}. \tag{2}$$

The clustering coefficient of the network is given by the average of C_i over all the vertices in the network:

$$C = \frac{1}{N}\sum_{i=1}^{N} C_i. \tag{3}$$

If a network has a high clustering coefficient C and a very short path length $\langle d\rangle$, it is a small world (SW) network [4]. In other words, a small-world graph can be defined as a network such that $\langle d\rangle \sim \langle d_{\text{rand}}\rangle$ and $C \gg C_{\text{rand}}$.

If the degree distribution of a network is following a power law:

$$P(k) \sim k^{-\gamma}. \tag{4}$$

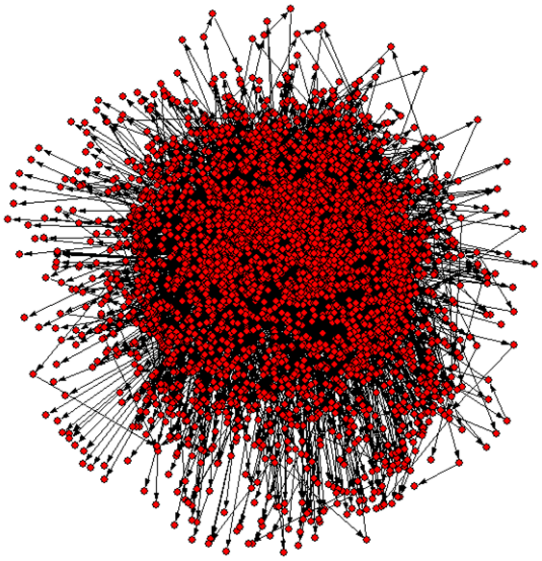

Fig. 4. Syntactic network built on xwlb treebank.

Table 2
Main properties of two syntactic dependency networks

Networks	N	$\langle k\rangle$	C	C_{rand}	$\langle d\rangle$	$\langle d_{\text{rand}}\rangle$	D	γ
xwlb	4017	6.48	0.128	0.00135	3.372	4.66	10	2.40
shss	2637	8.91	0.26	0.00362	2.996	3.83	10	2.18

N — the number of vertices, $\langle k\rangle$ — average degree, C — clustering coefficient, C_{rand} — clustering coefficient of random graph, $\langle d\rangle$ — average path length, $\langle d_{\text{rand}}\rangle$ — average path length of random graph, D — diameter, γ — exponent of power law.
C_{rand} and $\langle d_{\text{rand}}\rangle$ are obtained by calculating the E–R random network with the same parameters.

The network is a scale-free network, the constant γ is often between 2 and 3 [17]. In (4), $P(k)$ is the fraction of vertices in the network that have degree k. In other words, $P(k)$ is the probability that a vertex chosen uniformly at random has degree k.

Fig. 4 gives a panorama of xwlb syntactic dependency network. Fig. 3 can be seen as a fragment of a network such as Fig. 4.

We use Pajek and Minitab getting these properties as the following:

- The number of vertices (N): Info → Network → General.
- Average path length and diameter ($\langle d\rangle$ and D): Net → Paths between 2 vertices → Distribution of Distances → From All Vertices.
- Cluster coefficient (C): Net → Vector → Clustering Coefficients → CC1. To obtain distribution statistics for the CC, use "Info → Vector".
- Average degree ($\langle k\rangle$): Net → Partitions → Degree, that generates the Vector for the Degree Partition and use Info → Vector for the needed average value.
- Degree distribution and exponent of power law: Net → Partitions → Degree → All. Save the partition information as a file, that is imported into Minitab for getting a curve of degree distribution and exponent of power law with linear regression.

Table 2 and Fig. 5 show average path length, clustering coefficient and degree distribution of two syntactic dependency networks.

Fig. 5 shows that the degree distributions have noise fat-tails, which maybe make the non-perfect, but acceptable fitting with the power law. If we would neglect a few data points from Fig. 5, it is possible to get better fitness, with γ very close to 2. Another possible factor, which gives rise to the low exponents, is that our networks have fewer

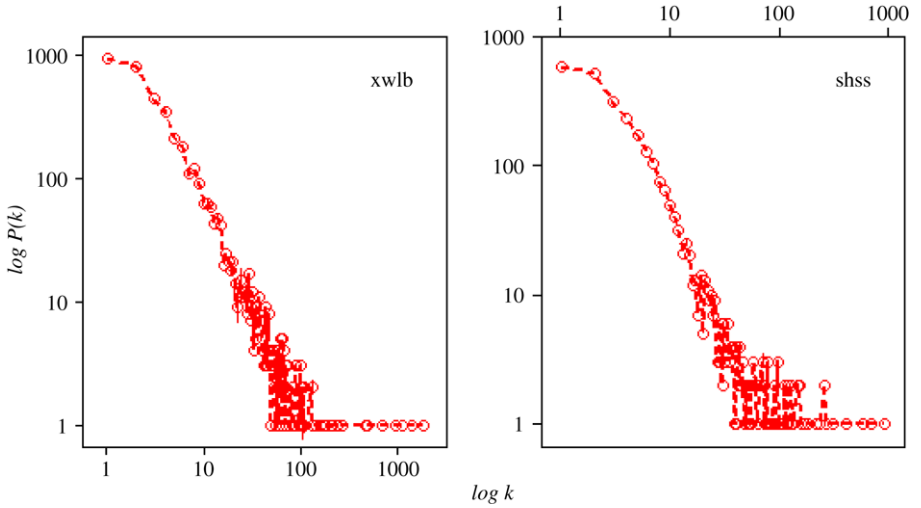

Fig. 5. Empirical degree distributions of two syntactic dependency networks. Two degree distributions were fitted by a power law with exponents 1.36 (xwlb) and 1.17 (shss) with $R^2 = 0.792$ and 0.791.

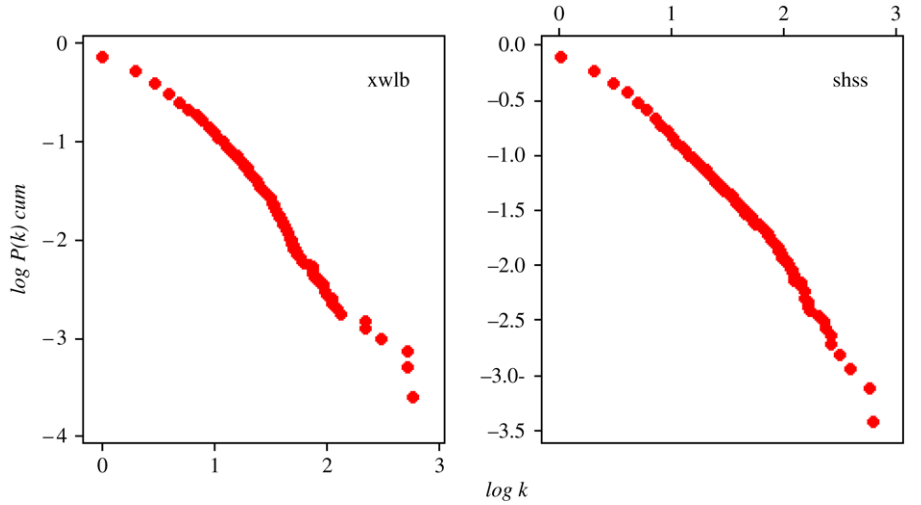

Fig. 6. Cumulative degree distributions of two syntactic dependency networks. Two cumulative degree distributions were fitted by a power law with slopes −1.40 (xwlb) and −1.18 (shss), which correspond to the exponents $\gamma = 2.40$ and 2.18 respectively.

vertices than other real network mentioned in Refs. [3,18], just as Dorogovtsev and Mendes point out that power law distributions are observable only in large networks [18, page 14].

For reducing the noise in the tail and making the exponents of power law more precise, we also make the cumulative degree distributions of two syntactic dependency networks in Fig. 6.

According to Ref. [3], the cumulative distribution function:

$$P(k) = \sum_{k'=k}^{\infty} P(k') \tag{5}$$

which is the probability that the degree is greater than or equal to k. If a degree distribution follows power law in its tail: $P(k) \sim k^{-\gamma}$ for some constant exponent γ. Then, such power-law distribution show up as power laws in the

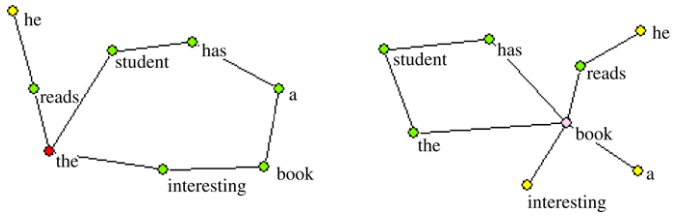

Fig. 7. Networks based on co-occurrence and syntactic principles. The left is co-occurrence and the right syntactic.

cumulative distributions too, but with exponent $\gamma - 1$ rather than γ:

$$P(k) \sim \sum_{k'=k}^{\infty} k'^{-\gamma} \sim k^{-(\gamma-1)}. \tag{6}$$

Fig. 6 presents the results of two cumulative degree distributions, which are well fitted by a power law with the determination coefficients R^2 0.969 and 0.967 respectively. By (5) and (6), we get more precise exponents of power law for two syntactic dependency networks, which are 2.40 and 2.18.

Compared with Figs. 5 and 6 presents two thicker tails that were generated due to some very common words being non-zero until large values, which maybe caused the disagreement of the exponents of degree distributions in the raw and cumulative forms. Ref. [19] reviews more than 20 studies on complex networks of linguistic units and concludes that "all these approaches should (but often fail to) answer the following questions:

(1) What do the vertices represent and subject to which criteria are they linked?
(2) What is the research interest in analyzing these networks?
(3) Which small-world or complex network indicators are investigated?
(4) Which reasons are assumed to evoke the small-world property if observed?
(5) Is there any account of network growth?"

It is time to answer and discuss these questions, several of which, such as (1)–(3), have been touched above. We proposed to build a syntactic network truly based on a linguistic (syntactic) theory and argued that a language network based on co-occurrence of words is not suitable for the goal of investigating the properties of a syntactic network of human language, although the analysis of co-occurrence network certainly has its informational-theoretic value [20]. Fig. 7 gives the networks of two English sentences in Table 1, which are built based on co-occurrence adjacent words and dependency syntax respectively.

Fig. 7 shows that two approaches build different networks, which would get different results from the same text. For instance, in the syntactic one, the noun "*book*" has the highest degree, but the determiner "*the*" plays that role in the co-occurrence network. Another factor, which may influence the cluster coefficient of the network, is the difference between the means to form clusters. While syntactic network never makes a link between an adjective and a determiner, it is normal to do that in co-occurrence network.

Our findings, according to which a syntactic network is small-world and scale-free, are similar to that of Ref. [21] based on Czech, Romanian and German treebanks. This reinforces the hypothesis formulated in the past by several authors about the universality of this network pattern in human language [7,19–21]. However, if we want to claim such similarities as one of the language universals, we still have to do further research, because our study also shows the similar properties between the two syntactic dependency networks and more than 30 real-world networks, which are not language related [3,18]. In the following discussions, we will try to link our empirical findings and linguistics.

Table 2 shows that the syntactic networks have similar average path length with E–R random networks, but the clustering coefficients of syntactic networks are much greater than that of E–R random networks. Therefore, the two syntactic networks are small-world networks.

Based on Romanian and Czech corpora, Ref. [22] investigates the Euclidean distance between syntactically linked words in sentences and finds that the average distance is significantly small and is a very slowly growing function of sentence length. Ref. [11] undertaken a similar experiment, but with the corpora of 20 languages, the results show that a threshold of average (dependency) distance is less than three words. The tendency of minimized average distance in a language may be one of the reasons that evoke the small-world property of syntactic networks.

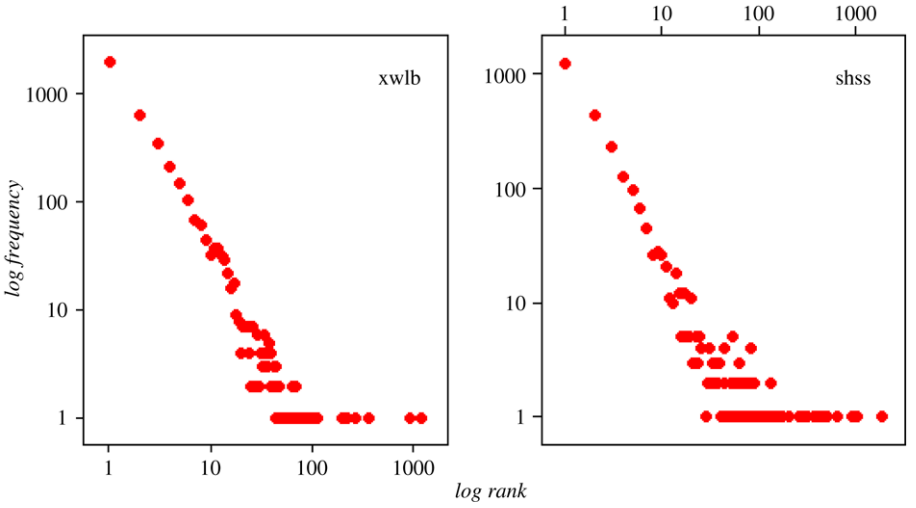

Fig. 8. Zipf curves of two treebanks. Regression lines with the slopes 1.278 (xwlb) and 0.912 (shss).

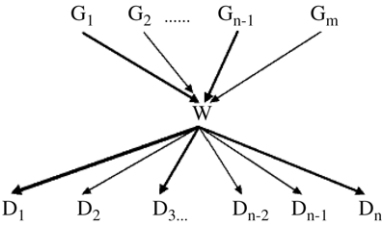

Fig. 9. Valency pattern of a word (class).

Figs. 5 and 6 show that the degree distributions of syntactic networks are power-law-like; this reveals that the two syntactic networks are scale-free-like.

Why does the degree distribution of a language network follow the power law? Ref. [20] argues that in a language network the degree of a word is equivalent to its frequency, whose distribution obeys the Zipf law [23] or power law. Thus, the degree distribution of a vertex also has such properties. Because a random text is also following the Zipf law [24,25], in this manner, we cannot explain the role of syntax in a language network. Fig. 8 shows Zipf curves of two treebanks, which are also sources of our syntactic networks.

Comparing Figs. 5 and 8, we can find that the degree distributions of two syntactic networks are not equivalent to their Zipf curves, and the regression lines have different slopes, although they all are similar and scale-free-like. Thus, the syntactic factor may still play a role during the forming of the degree distribution of a syntactic network. To investigate the difference, we extract top 15 words according to degree and frequency from the networks and treebanks. The results are in Table 3.

Table 3 reveals that the degree of a word is not equivalent to its frequency, although both are close related. In the lists of top 15 words, functional or grammatical words play an important role in degree and rank by frequency. It is worthwhile to notice the difference regarding content words: while in the frequency list, noun is the authority, in the degree list, a verb takes the leading role. The empirical finding links the network with dependency syntax, which essentially is a theory based on valency of a word and the verb is the center of a sentence [26]. Ref. [27] develops the idea in Ref. [26] and proposes a new syntactic theory "Probability Valency Pattern", which claims that almost all words (linguistic units) have a potential capacity for combining with other words to form bigger linguistic units. The potential capacity can be called as the valency (pattern) of the word. Valency contains centripetal (input) and centrifugal (output) forces. While the centripetal force is the capacity to be governed by other words, centrifugal measures the capacity to govern other words. The valency pattern of a word can be sketched as in Fig. 9.

Here W is a word or word class. $G_1, G_2, \ldots, G_m$ are different dependency types, which be able to govern W. $D_1, D_2, \ldots, D_n$ are dependency types possibly governed by W. When the word comes into the text, the potential capacity

Table 3
The degree and frequency of top 15 words in the networks and treebanks

xwlb				shss			
Degree	Word	Word	Frequency	Degree	Word	Word	Frequency
935	的	的	930	833	的	的	1061
525	和	和	276	614	是	我	946
227	在	在	228	387	有	是	638
225	是	了	202	310	我	个	517
135	了	是	114	264	个	一	494
126	为	发展	106	234	了	了	430
114	有	中国	104	228	说	这	426
113	要	一	94	204	在	们	374
100	对	经济	91	165	看	不	322
98	个	对	85	161	要	他	295
94	进行	中	83	151	想	你	288
87	说	工作	80	150	到	有	277
84	发展	为	75	142	觉得	就	260
80	与	个	70	139	能	说	205
77	到	将	69	123	人	在	177

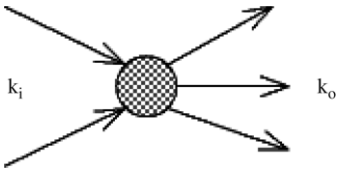

Fig. 10. An abstract valency of a word. k_i and k_o are the govern and governed valency of a word.

is activated and begin to combine with other words to generate a dependency analysis of a sentence. If we see the word as a vertex in the network, and G and D as the in- and out-degrees, Fig. 9 can be abstracted into Fig. 10.

Fig. 10 is almost same as the description of a vertex in network science [18:6, Fig. 1.2]. In this manner, we build a theoretical link between syntactic theory and network science, and find some linguistic roots of syntactic network with small-world and scale-free properties, but more mathematical models still have to been built to better understand the essence and dynamics of the syntactic network. It is one of the directions of further research.

In this paper, we used two genres, which differ greatly, to construct syntactic dependency networks in order to find if such differences can also be found by network's analysis in particular, and if the complex network can be used as a means to investigate linguistic questions in general. A few studies have been done on the possible applications of complexity theory in linguistics, for instance, Ref. [28] proposes an approach to study the syntactic development using complexity theory. The discussions above show that it is possible to explore syntactic networks from valency theory and dependency grammar, or conversely.

Table 2 shows that the two networks have same diameters, but their average degree, average path length, clustering coefficients and exponents of power law are different.

4. Concluding remarks

The researchers, who study the language networks from modern network science, often ignore the achievements of linguistics. Thus, the conclusions of network science are difficult to explain from a linguistic point of view. This makes the study of the dynamics of language networks less useful and it reduces the importance of these

studies for developing linguistics. Therefore, we propose that language networks should have linguistic theory as its fundamentals.

The paper reports how to build a syntactic network based on dependency treebank, and analyzes two Chinese syntactic dependency networks using the techniques from complex networks. The results show that the two syntactic networks are small-world and scale-free-like.

We try to explain the findings from linguistics and linked the degree of a vertex with a valency of a word in syntactic theory, which may be useful to study the dynamics of a syntactic network.

The paper compares the network indicators of two syntactic networks with different text genres. They have similar diameters, but different average degree, path length, power exponents and clustering coefficients. It seems possible to use complexity as a means of stylistic study.

Further studies planned to elucidate remaining issues which can also be seen as pointing to the potential problems and limitations of the current approach, are:

(1) how to process some unusual syntactic structures in dependency syntax, e.g., coordinating structures;
(2) how to find the network indicators of a directed syntactic network and link these findings with linguistic theories;
(3) how to convert the current network to a weighted syntactic network, aiming at exploring the link strength in a syntactic network.

These works will make our current network a directed and weighted network, which seems to be a more fitting model for a language network based on dependency syntax.

In summary, for making more convincing conclusions and interesting findings, we still need to conduct further studies with much more materials and genres.

Acknowledgments

We thank Anat Ninio for detailed comments, the reviewers for insightful comments, Zhao Yiyi and Guan Runchi for annotating treebanks.

References

[1] S.M. Lamb, Pathways of the Brain: The Neurocognitive Basis of Language, John Benjamins, Amsterdam, 1998.
[2] R.A. Hudson, Language Networks: The New Word Grammar, Oxford University Press, Oxford, 2007.
[3] M.E.J. Newman, The structure and function of complex networks, SIAM Review (2) (2003) 167–256.
[4] D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, Nature 393 (1998) 440–442.
[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: Structure and dynamics, Physics Reports 424 (2006) 175–308.
[6] R. Solé, B. Corominas, S. Valverde, L. Steels, Language networks: Their structure, function and evolution, Santa Fe Institute Working Paper (05-12-042), 2005.
[7] R. Ferrer i Cancho, The structure of syntactic dependency networks: Insights from recent advances in network theory, in: G. Altmann, V. Levickij, V. Perebyinis (Eds.), The Problems of Quantitative Linguistics, Ruta, Chernivtsi, 2005, pp. 60–75.
[8] Y. Li, L.X. Wei, et al., Small-world patterns in Chinese phrase networks, Chinese Science Bulletin 50 (3) (2005) 286–288.
[9] Y. Li, L. Wei, W. Li, Y. Niu, S. Luo, Structural organization and scal-free properties in Chinese phrase networks, Chinese Science Bulletin 50 (13) (2005) 1304–1308.
[10] J. Li, J. Zhou, Chinese character structure analysis based on complex networks, Physica A 380 (2007) 629–638.
[11] H. Liu, Dependency distance as a metrics of comprehension difficulty (2007) (submitted for publication).
[12] J.-Y. Ke, Y. Yao, Analyzing language development from a network approach, Journal of Quantitative Linguistics 15 (1) (2008) 70–99.
[13] J. Nivre, Inductive Dependency Parsing, Springer, Dordrecht, 2006.
[14] A. Abeillé (Ed.), Treebank: Building and using Parsed Corpora, Kluwer, Dordrecht, 2003.
[15] W. de Nooy, A. Mrvar, V. Batagelj, Exploratory Social Network Analysis with Pajek, Cambridge University Press, Cambridge, 2005.
[16] B.F. Ryan, Brian L. Joiner, Jonathan D. Cryer, MINITAB Handbook: updated for Release 14, 5th edition, Duxbury Press, Belmont, 2005.
[17] A-L Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[18] S.N. Dorogovtsev, J.F.F. Mendes, Evolution of Network: From Biological Nets to the Internet and WWW, Oxford University Press., London, 2003.
[19] A. Mehler, Large text networks as an object of corpus linguistic studies, in: Anke Lüdeling, Kytö Merja (Eds.), Corpus Linguistics. An International Handbook of the Science of Language and Society, De Gruyter, Berlin, New York, 2007.
[20] A.P. Masucci, G.J. Rodgers, Network properties of written human language, Physical Review E 74 (2006) 026102.
[21] R. Ferrer i Cancho, R.V. Solé, R. Köhler, Patterns in syntactic dependency networks, Physical Review E 69 (2004) 051915.
[22] R. Ferrer i Cancho, The Euclidean distance between syntactically linked words, Physical Review E 70 (2004) 056135.

[23] G.K. Zipf, Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology, Addison-Wesley Press, Cambridge, MA, 1949.
[24] W. Li, Random texts exhibit Zipf's-law-like word frequency distribution, IEEE Transactions on Information Theory 38 (6) (1992) 1842–1845.
[25] R. Ferrer i Cancho, Ricard V. Solé, Zipf's law and random texts, Advances in Complex Systems 5 (2002) 1–6.
[26] L. Tesnière, Eléments de la syntaxe structurale, Klincksieck, Paris, 1959.
[27] H. Liu, Zh. Feng, Probabilistic valency pattern theory for natural language processing, Language Science 3 (2007) 32–41 (in Chinese).
[28] A. Ninio, Language and the Learning Curve: A New Theory of Syntactic Development, Oxford University Press, London, 2006.