Extracting Valency Patterns of Word Classes from Syntactic Complex Networks

Chen Xinying Xu Chunshan Li Wenwen
Communication University of China, Beijing
cici13306@gmail.com

Abstract

Our study extracted different valency patterns of Chinese word classes from 3 different Chinese dependency syntactic networks and compared their similarities and differences. The advantages and disadvantages of network approach are discussed at the end of this paper. The results show that there are some persisting properties in Chinese which are not affected by style. There are also some word classes which are more sensitive to the stylistic impact in Chinese. The network approach to linguistic study can make the complex data concise and easy to understand. However, it also has some deficiencies. First of all, when the network size is large, the structure will become so complex that easy understanding is impossible. Secondly, although the network can easily provide an overview of the language, it usually fails to be much helpful when it comes to language details.

Introduction

Reductionism has driven 20th century science, with the result being that we have experts who know more and more about less and less while leaving us devoid of generalists and multidisciplinary artists and scientists who can "connect the dots" across these fragmented foci（Albert-László Barabási 2002）. Now, more and more people realize that we can not figure out the overall structure by researching parts. In this situation, the network science, which provides a way to study the relationship between parts from an overall perspective, has a rapid development.

Language system is a complex network (Hudson, 2007). Therefore, the use of complex networks is a necessary attempt to study language (Liu, 2011). The research of language network can give a global perspective about language structure and about the relationship between language units.

There have been many researches on language complex networks (Liu, 2010; Ferrer i Cancho *et al*, 2004; Yu, 2011). Although the networks are built at different levels of language and with different concerns, most studies put the emphasis on the common features of various networks, such as small world and scale-free characteristics. This research approach is novel, and the results are often difficult to interpret in terms of linguistic theories. It seems that the study of language network lacks a solid foundation of linguistic theories. For linguists, network is simply a means and a tool for linguistic study but not the goal. We hope to establish a close link between the network and linguistic theories and study how network can serve to local syntactic studies or semantic studies, so the network can play a more important role in linguistic study. The paper tries to making an insightful exploration in this direction.

In language networks, such as syntactic and semantic networks, the nodes are language units on the same linguistic level which have direct relationships with one another. From this perspective, the Valence Theory is rather network-friendly, which provides a suitable linguistic theory to explain the findings of studies of language networks.

The theoretical basis of our study is Probabilistic Valency Pattern (PVP, Liu 2006), which is developed from the classic Valence Theory. The study extracted 3 different valency patterns of Chinese word classes from 3 different Chinese dependency syntactic networks and compared their similarities and differences. The discussion about the advantages and disadvantages of the network approach on linguistic study will be presented at the end of the paper.

PVP

The Valence Theory has been developed and revised in many ways since Tesnière (1959) integrated valence into syntactic theory. The

concept of valency can be found in almost all modern linguistic theories.

Traditionally, the Valence Theory is a syntactic-semantic theory. It is a term used to describe the relationship between a language unit and its complement. With the development of computer technology, people began to use computers to analyze and process real language. To analyze real language, taking only the complement into account is not enough. Under such circumstances, Liu (2006) propose PVP. After surveying the definitions of valence claimed by Tesnière (1959), Helbig (1978 2002), Schenkel (1978), Fischer (1997), Mel'čuk (2003), Hudson (2004) and others, Liu found that, despite some differences among these definitions, one thing remains unvarying: valence is the combinatorial capacity of a word. They argued that this is the general definition of valence: the combinatorial capacity shared by all words as one of their fundamental attributes. The capacity is a potential one whose realization is constrained by syntactic, semantic and pragmatic factors. When a word comes into a text, the potential capacity is activated, producing a dependency relation and establishing a pattern of sentence structure. They also put forward that the dependencies involved in the valence of a word may distribute unevenly. The probability information should be added to the descriptions of words' valences so as to indicate the strength of corresponding combinations. According to PVP, it is necessary to qualitatively describe the dependencies involved in the valence of a word or word class.

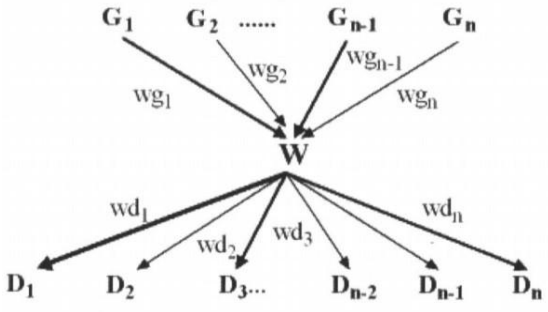

Figure 1. Valency Patterns in PVP (Liu 2006)

The 'W' can be a word class or a specific word. G1, G2 ... Gn are dependencies that take 'W' as the dependent. D1, D2 ... Dm are dependencies that take 'W' as the governor. wg1, wg2 ... wgn are the probabilities of different dependencies and wg1 + wg2 + ... + wgn = 1. It is the same with wd1, wd2 ... Wdm.

Extracting valency patterns of Chinese word classes from syntactic complex networks

PVP represents the probabilistic dependencies between words or word classes. Before the exploration into language details, an overview of Chinese dependency structure will be of much help. So we chose to study the valency patterns of word classes instead of words. There are no similar researches so far in literature. First, we built 3 Chinese dependency syntax treebanks and then converted them into dependency syntactic networks. After that, from these networks we extracted the valency patterns of Chinese word classes and compared their similarities and differences. Treebanks are the basis of this study. Taking stylistic influences into account, we selected the "实话实说" shi-hua-shi-shuo 'name of a famous Chinese talk show' (hereinafter referred to as SHSS) and "新闻联播" xin-wen-lian-bo 'name of a Chinese TV news program' (hereinafter referred to as XWLB), two corpora with different styles, for annotation. We transcribed and annotated these two oral corpus. The annotation scheme is the Chinese Dependency Annotation System proposed by Liu (2006). SHSS is colloquial, containing 19, 963 words. XWLB is of a quite formal style, containing 17,061 words. In order to get a corpus which can reflect the general structure of Chinese without the reflections of language styles, we put the SHSS and XWLB together and get the third corpus. We respectively built 3 treebanks with SHSS, XWLB and SHSS+XWLB (hereinafter referred to as the S-treebank, X-treebank, A-treebank). Table 1 shows the format of our Chinese dependency treebanks.

This format includes all the three mentioned elements of the dependency relation, and can easily be converted into a graph as shown in Figure 2.

Order number of sentence	Dependent			Governor			Dependency type
	Order number	Character	POS	Order number	Character	POS	
S1	1	这	r	2	是	v	subj

S1	2	是	v	6	。	bjd	s
S1	3	一	m	4	个	q	qc
S1	4	个	q	5	苹果	n	atr
S1	5	苹果	n	2	是	v	obj
S1	6	。	bjd				

Table 1. Annotation of a sample sentence in the Treebank[1]

[1] The details of all codes and symbols in tables and figures in this paper are available in Appendix A.

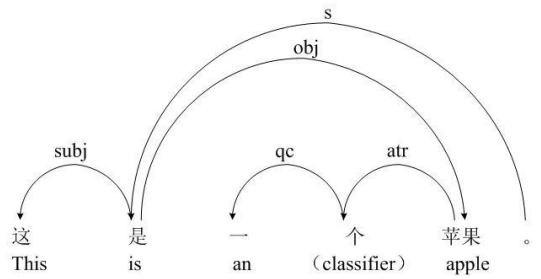

Figure 2. The graph of the dependency analysis of a sentence

With words as nodes, dependencies as arcs and number of dependencies as the value of arcs, networks are built, in which the direction of arc is defined as from governor nodes to dependent nodes. For example, the sample shown in Figure 2 can be converted to a network structure as shown in Figure 3 (excluding punctuation). Figure 4 presents the syntactic network converted from A-treebank.

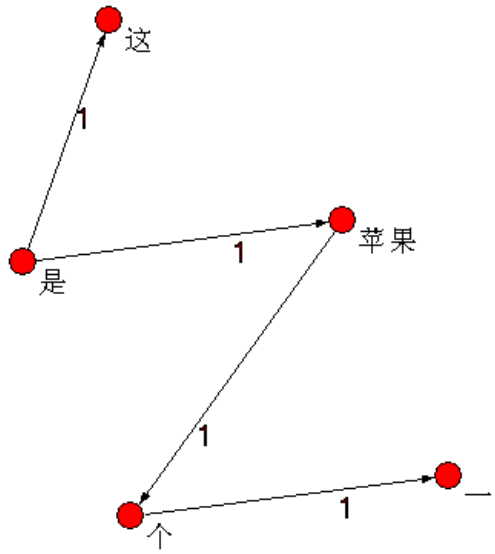

Figure 3. Network of 这是一个苹果 zhe-shi-yi-ge-ping-guo ‘this is an apple’

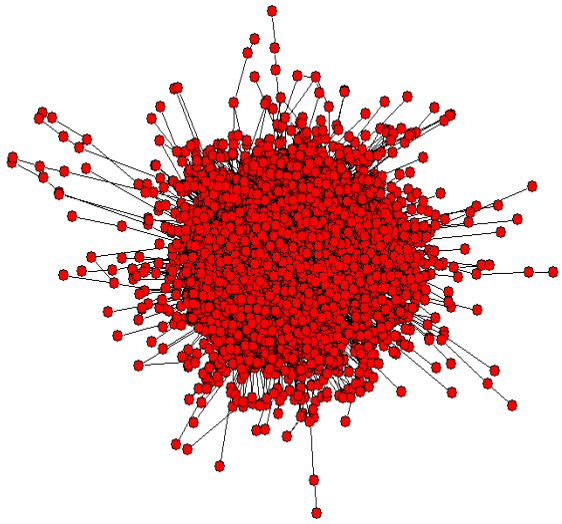

Figure 4. Network of A-treebank

In Figure 4, the nodes are words. We can also cluster the nodes which belong to the same word class into a new node and the new node will inherit all the dependencies of these nodes. In this way, we can obtain a dependency network of word classes. We extracted 3 networks of word classes from the S-treebank, X-treebank and A-treebank. For the sake of clarity, the values of arc, which are given in Table 2, 3 and 4, are not shown in Figure 5, 6, 7.

In Table 2, the first column is the list of dependent word classes and the first row is the list of governing word classes. In each cell in this table is the frequency of the dependency relation between a certain governing word class and a certain dependent governing word class, or, the value of the corresponding arc.

Governor / Dependent	d	r	m	a	u	c	p	q	v	n	o	e	zmen	zdi
d	9	0	4	103	8	7	7	1	567	17	0	0	0	0
r	2	5	8	6	35	6	18	39	154	164	0	0	0	0
m	3	0	80	7	11	5	5	396	84	160	0	0	0	0
a	1	1	7	46	135	29	8	8	384	422	0	0	0	0
u	0	1	17	20	2	58	3	6	415	749	0	0	0	0
c	1	1	1	7	44	8	24	0	293	83	0	0	0	0
p	3	1	0	5	56	6	3	0	638	13	0	0	0	0
q	1	0	2	4	13	5	10	6	121	274	0	0	0	0
v	5	4	15	23	365	289	119	8	2216	635	0	0	0	0
n	4	13	15	105	349	474	527	24	3046	2558	0	0	0	0
o	0	0	0	0	0	0	0	0	0	0	0	0	0	0
e	0	0	0	0	0	0	0	0	0	0	0	0	0	0
zmen	0	0	0	0	0	0	0	0	0	4	0	0	0	0
zdi	0	0	9	0	0	0	0	0	0	0	0	0	0	0

Table 2. values of arcs (X-treebank)

Governor / Dependent	d	r	m	a	u	c	p	q	v	n	o	e	zmen	zdi
d	7	6	9	341	26	3	16	3	1481	13	0	0	0	0
r	2	62	19	93	229	14	136	432	1652	325	0	0	0	0
m	0	1	64	15	6	0	0	603	40	79	0	0	0	0
a	1	2	5	55	240	6	7	7	321	220	0	0	0	0
u	10	10	2	97	16	9	8	9	1090	668	0	0	0	0
c	0	6	0	28	7	3	10	2	338	32	0	0	0	0
p	0	1	0	26	18	3	1	0	460	10	0	0	0	0
q	2	15	3	26	19	1	6	10	278	692	0	0	0	0
v	11	8	2	124	315	22	44	7	3484	187	0	0	0	0
n	5	58	3	141	198	69	302	7	2382	688	0	0	0	0
o	0	0	0	0	0	0	0	0	1	0	0	0	0	0
e	0	0	0	2	0	0	0	0	1	0	0	0	0	0
zmen	0	361	0	0	0	0	0	0	1	8	0	0	0	0
zdi	0	0	31	0	0	0	0	0	0	0	0	0	0	0

Table 3. values of arcs (S-treebank)

Governor / Dependent	d	r	m	a	u	c	p	q	v	n	o	e	zmen	zdi
d	16	6	13	444	34	10	23	4	2048	30	0	0	0	0
r	4	67	27	99	264	20	154	471	1806	489	0	0	0	0
m	3	1	144	22	17	5	5	999	124	239	0	0	0	0
a	2	3	12	101	375	35	15	15	705	642	0	0	0	0
u	10	11	19	117	18	67	11	15	1505	1417	0	0	0	0
c	1	7	1	35	51	11	34	2	631	115	0	0	0	0
p	3	2	0	31	74	9	4	0	1098	23	0	0	0	0
q	3	15	5	30	32	6	16	16	398	966	0	0	0	0
v	16	12	17	147	680	311	163	15	5701	822	0	0	0	0
n	9	71	18	246	547	543	829	31	5428	3246	0	0	0	0
o	0	0	0	0	0	0	0	0	1	0	0	0	0	0
e	0	0	0	2	0	0	0	0	1	0	0	0	0	0
zmen	0	361	0	0	0	0	0	0	1	12	0	0	0	0
zdi	0	0	40	0	0	0	0	0	1	0	0	0	0	0

Table 4. values of arcs (A-treebank)

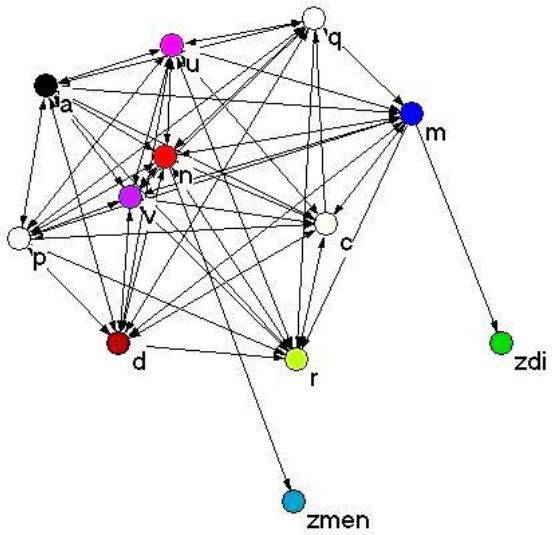

Figure 5. Network of word classes (X-treebank)

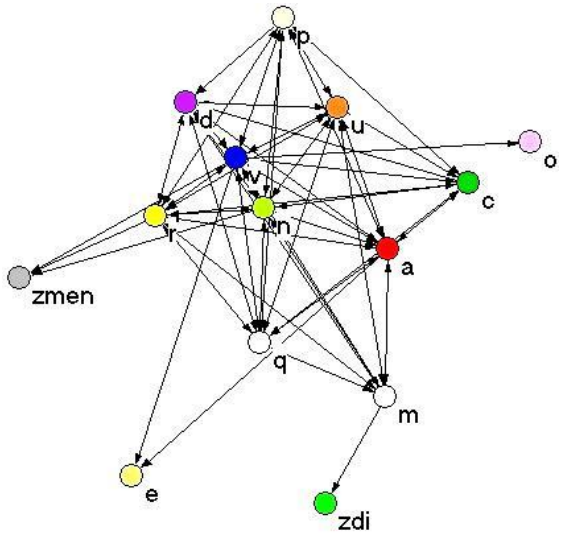

Figure 6. Network of word classes (S-treebank)

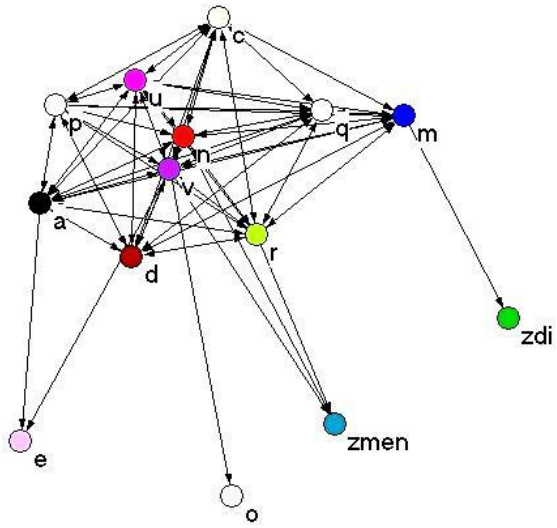

Figure 7. Network of word classes (A-treebank)

Since the PVP describes the combinatorial strength between words or word classes, the frequencies of dependencies, which reflect the actual realization of valence, can be used to quantitatively measure the combinatorial strength. Because the values of arcs indicate frequencies of dependencies between the nodes, the greater the value is, the greater this strength is in network.

Figure 1 can be seen as the valency pattern of an unknown word class. Obviously, the valence patterns of all word classes will merge into a complex network whose nodes are word classes, that is, a dependency networks as shown in Figure 5, 6 and 7. The network structure can be understood as roughly reflecting valency patterns, the nodes representing word classes and the arcs representing dependencies. The direction of arcs distinguishes governing nodes from dependent nodes and the values of arcs indicate the frequencies of dependencies, from which the probability can easily derive. We use, instead of numbers, the distance between nodes to indicate the combinatorial strength between network nodes, which, we hope, can make for an easy understanding of the combinatorial strength between nodes. The higher the value of arc, the greater the combinatorial strength and the shorter the distance.

Results and discussion

Based on these 3 networks, we found that the valency patterns of X-treebank, S-treebank and A-treebank have a few things in common:

(1) 'zdi', 'zmen', mimetic word and interjection cannot be governors.
(2) 'zdi' can only be the dependent of numeral.
(3) mimetic word can only be the dependent of verb.
(4) The distance between the verb and noun is the shortest. In other words, the dependency between noun and verb has the highest probability in all dependencies between different word classes.
(5) The governor of adjective is most likely to be verb. The dependent of adjective is most likely to be adverb.
(6) The dependent of pronoun is most likely to be 'zmen'.
(7) The governor of numeral is most likely to be classifier. The dependent of numeral is most likely to be numeral.
(8) The two most probable governors of classifier are noun and verb. The dependent of classifier is most likely to be numeral.
(9) The governor of preposition is most likely to be verb. The dependent of preposition is most likely to be noun.
(10) The dependent of auxiliary is most likely to be verb.
(11) The governor of adverb is most likely to be verb.
(12) The governor of conjunction is most likely to be verb. The dependent of conjunction is most likely to be noun.

These phenomena are found in all 3 networks which mean that they are unaffected by stylistic impact and are stable properties of this language. When we process the real language, it is a good choice to give these properties the priority, which can promote the efficiency avoiding random analysis.

We have also found several dissimilarities:

(1) There are no mimetic word and interjection in valency pattern of X-treebank.
(2) The value of arcs invoving pronoun is the smallest in X-treebank while it is largest in S-treebank. It is means that pronoun is sensitive to language styles.
(3) In X-treebank, the probability of auxiliary linking with noun is higher than that of auxiliary linking with verb. It is opposite to the valency pattern of S-treebank. It is also may be seen as the different property between language styles.

These results prove that some word classes could show different valence patterns in texts with different styles. If we want to describe the valency pattern accurately, we should find a way to reduce the stylistic impact. So A-treebank may present a more reliable valence pattern of word classes. From the data, we can

see that mimetic word, interjection, pronoun and auxiliary are more sensitive to the stylistic impact. It shows that network approach and PVP may be able to provide some effective parameters for Chinese style research.

Conclusion

With three dependency networks, we have found that the valence pattern can be affected by style; simultaneously we investigated the similarities and differences among them. This work is trying to study the valence patterns of the language from an overall perspective and compare different valence patterns then figure out the real structure of language systems. It is different from traditional statistical works on words or word classes, for example collocation extraction from tree banks etc., which pay more attention to some specific structures.

In the study, we found that the language network approach and PVP are beneficial to each other. PVP can explain the language network data, such as the node, arc, value of arc, direction of arc, distance between nodes, etc. At the same time, as a method of language study, complex network can provide an intuitionistic but concise representation of data, which is easy to perceive and understand. However network approach also has some deficiencies. First of all, when the network size is large, the structure will become so complex that easy understanding is impossible. Secondly, although the network can easily provide an overview of the language, it usually fails to be much helpful when it comes into language details. For example, we cannot give the arcs qualitative descriptions in the network, which implies the loss of valuable information. Extracting valency patterns of word classes from syntactic complex networks is an interesting experiment. The integration of language network and PVP makes us believe that further research will bring more valuable results.

References

Albert-László Barabási. 2002. Linked. Cambridge, Perseus Publishing.

Ferrer i Cancho, R., R. Koehler and R. V. Solé 2004. Patterns in Syntactic Dependency Networks. *Physical Review E*, 69.

Fischer, K. 1997. *German-English Verb Valency*. Tübingen, Gunter Narr Verlag.

Helbig, G. and Schenkel, W. 1978. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig, Biblio graphishes Institut.

Helbig, G. 2002. *Linguistische Theorien der Moderne*. Berlin, Weidler Buchverlag.

Hudson, R. 2004. *An Encyclopedia of English Grammar and Word Grammar*. Http://www.phon.ucl.ac.uk/home/dick/encgen.htm.

Hudson, R. 2007. *Language Networks: The New Word Grammar*. Oxford, Oxford University Press.

Liu, H. 2006. Syntactic Parsing Based on Dependency Relations. *Grkg/Humankybernetik*, 47:124-135.

Liu, H. 2010. Language Clusters based on Linguistic Complex Networks. *Chinese Science Bulletin*, 55.

Liu, H. 2011. Linguistic Networks：Metaphor or Tool? *Journal of Zhejiang University (Humanities and Social Science)*. 41: 169-180.

Liu, H. and Huang, W. 2006. A Chinese Dependency Syntax for Treebanking. *Proceedings of The 20th Pacific Asia Conference on Language, Information and Computation*. Beijing, Tsinghua University Press, 126-133.

Mel'čuk, I. 2003. Levels of Dependency in Linguistic Description: Concepts and Problems. In V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, H. Lobin (eds): *Dependency and Valency. An International Handbook of Contemporary Research*, vol. 1. Berlin - New York: W. de Gruyter, 188-229.

Tesnière, L. 1959. *Éléments de syntaxe structurale*. Paris, Klincksieck.

Yu, S., Liu, H. and Xu, C. 2011. Statistical Properties of Chinese Phonemic Networks. *Physica A*, 390.

Appendix A. codes meaning

code	meaning
d	adverb
r	pronoun
m	numeral
a	adjective
u	auxiliary
c	conjunction
p	preposition

q	classifier
v	verb
n	noun
o	mimetic word
e	interjection
zmen	“们” men ‘kind of suffix’
zdi	“第” di ‘kind of prefix’
bjd	punctuation
subj	subject
s	main governor
qc	complement of classifier
atr	attributer
obj	object