Dependency Network Syntax

From Dependency Treebanks to a Classification of Chinese Function Words

Xinying Chen
School of International Study
Xi'an Jiaotong University, China
chenxinying@mail.xjtu.edu.cn

Abstract

This article presents a new approach of using dependency treebanks in theoretical syntactic research: the view of dependency treebanks as combined networks. This allows the usage of advanced tools for network analysis that quite easily provide novel insight into the syntactic structure of language. As an example of this approach, we will show how the network approach can provide clear structural distinctions among the Chinese function words, which are very difficult to obtain directly from the original treebank. We hope to illustrate the enormous potential of the language network approach through a simple example.

1 Why treebanks?

Treebanks are the latest hype in linguistics. The interest in treebanks can roughly be explained by two main charms: the NLP push to data driven approaches and the linguist's fascination of creating a treebank following specific theoretical principles.

In greater detail, we can observe that Natural Language Processing requires treebanks for all kinds of data-driven approaches ranging from Machine Translation to text classification. Great efforts, monetary and personal, go into the creation of treebanks or the transformation of existing treebanks into new formats. In particular dependency treebanks offer interesting connections between texts and the representation of meaning, the ultimate goal of Computational Linguistics. This NLP interest in dependency treebanks has also enthused (and frequently financed) the community of "pure" linguists, who have discovered that the creation of coherent treebanks is linguistically challenging and fascinating. Work has been done on error detection (Dickinson & Meurers 2003), alignment of multilingual (Lopez et al. 2002) and of multi-stratal treebanks (Böhmová et al. 2003; Mille & Wanner 2010), on written and spoken data, just to name a few. The creation of a treebank can also have a unifying effect on a linguistic community by providing a reference analysis, other analysis have to be compared to (Penn Tree Bank, Marcus et al. 1993). But the creation of a treebank following a specific syntactic theory cannot in itself be considered as a confirmation of this theory (other than being a sociological proof of the existence of sufficient support for the theory to be able to create a treebank).

What is crucially missing in this picture is the usage of treebanks for linguistic discovery and theory confirmation or refutation that goes beyond searching for examples in the annotated data. Simple concordancers exist, some of them with sophisticated query languages (Zeldes et al. 2009) but it is up to the syntactician to go through the results and make conclusions. No generally accepted approach on how to interpret this type of data has been established.

The community of "corpus linguistics" is nearly exclusively busy with statistical analysis on pure text corpora, using tools like Wordsmith or Lexico3. At most, they use POS tagged corpora, often simply to disambiguate word usages. This domain of research has achieved impressive results in historical linguistics, sociolinguistics, and other domains where large amounts of data finally are systematically ploughed through (Baker 1993; Charteris-Black 2004). However, the mentioned tools and methods can not easily be applied to treebanks because first, the structure of the data is very different, and secondly, the limited size of treebanks compared to the vast amounts of unannotated text, makes a statistical approach less interesting.

Notable exceptions to this rule include work on usually small hand-coded treebanks like the ones used in Liu et al. (2009) for the study of dependency distance and in Liu (2009a) for the research of probability distribution of dependencies, where the traditional statistical approaches have shown their potential in theoretic syntactic research. As an emerging statistical method, the network approach brings a new angle to this type of research.

In this paper, we attempt to illuminate the network view of dependency treebanks. We will show how this approach reduces the diffi-

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 41–50,
Prague, August 27–30, 2013. © 2013 Charles University in Prague, Matfyzpress, Prague, Czech Republic

culties in exploiting treebank data and how this approach can be successfully applied to small dependency treebanks, reducing the size limitations of existing dependency treebanks.

2 Language networks

The basic idea underlying dependency networks is very simple: instead of viewing the trees as linearly aligned on the sentences of the corpus, we fuse together each occurrence of the same word to a unique node, thus creating a unique and (commonly) connected network of words, in which the tokens are the vertices and dependency relations are the edges or arcs. This connected network is then ready to undergo common network analysis with tools like UCINET (Borgatti et al. 2002), PAJEK (Nooy et al. 2005), NETDRAW (Borgatti 2002), CYTOSCAPE (Shannon 2003), and so on.

In reality, extracting a network from a dependency treebank is slightly more complicated, as we have to use some heuristics to fuse together only the words that belong to the same lexeme (same category, near meaning). We refer to Liu (2008) for a description of multiple ways of network creation from dependency treebanks.

Linguistic research with using modern network analysis tools is an upcoming domain. The first conference on this subject, Modeling Linguistic Networks, was held in December 2012 in Frankfurt and united nearly 40 scholars from 14 countries. This community is guided by two assumptions: First, Language is physiognomicly a network and modeling of language should follow this guiding principle, and secondly, computational tools that have proven to be successful in sociology and computer science can be used for language networks, too.

The key interest of the network approach in linguistic research is that it provides a new way to analyze language systems. A central assumption of modern linguistic theories is that language is a system (Kretzschmar 2009). This widely accepted point of view, however, has remained on a purely theoretic level due to the absence of an operational methodology, until corpora and modern network analysis tools appeared. As language is a system, we expect there to be rules that cannot be predicted directly on the basis of the units. So looking at some specific words (or the relationship between them) may not be an efficient way for discovering the global features of a language system. Modeling language as a network provides an operational way for observing the macroscopic features of language system and the relationship between the units and the whole system. For example, it can be used for determining the function or status of some units, such as words, in the language system as a whole.

Some research has been done on the structure of syntactic dependency networks (Ferrer i Cancho 2005; Liu 2008; Chen & Liu 2011; Čech et al. 2011), the patterns in syntactic dependency networks (Ferrer i Cancho 2004; Chen et al. 2011), the language development or language evolution (Ke & Yao 2008; Mukherjee et al. 2013; Mehler et al. 2011), language clustering and linguistic categorization (Liu 2010; Liu & Cong 2013; Gong et al. 2012; Abramov & Mehler 2011), manual and machine translation (Amancio et al. 2008 &2011), word sense disambiguation (Christiano Silva & Raphael Amancio 2013), communication and interaction (Banisch et al. 2010; Mehler et al. 2010), the structure of semantic networks (Borge Holthoefer & Arenas 2010; Liu 2009b), phonetics (Arbesman et al. 2010; Yu et al. 2010), morphology (Čech & Mačutek 2009; Liu & Xu 2011), parts of speech (Ferrer i Cancho et al. 2007), Knowledge Networks (Allee 2000), cognitive networks (Mehler et al. 2012).

Works on Chinese include networks that use as nodes the Chinese characters (Li & Zhou 2007; Peng et al. 2008), words and phrases (Li et al. 2005), phoneme and syllables (Yu et al. 2011; Peng et al. 2008), syntactic structure (Liu 2008; Liu 2010; Chen & Liu 2011; Chen et al. 2011), semantic structure (Liu 2009b).

In general, the language network research, including that on Chinese language network, is developing rapidly in recent years. But the language network research inevitably has some aspects that need to be improved in order to establish this new domain. It seems that most of the language networks studies put a heavy emphasis on common features of various networks, such as ‘small world’ (Watts & Strogatz 1998) and ‘scale-free’ (Barab ási & Bonabeau 2003) features, treating alike different levels of language and different concerns on which the networks are built. At the same time, many language networks were built without proper guide of a specific linguistic theory, such as words’, characters’, or phrases’ co-occurrence networks (Li & Zhou 2007; Peng et al. 2008; Liu & Sun 2007; Li et al. 2005), resulting in research that lacks a strong connection to existing linguistic theories and research. But as more and more linguists get involved in the study of language networks, this situation is gradually changing.

3 The Chinese Dependency Network for this study

For the present work, we used the following treebank of Chinese: The treebank has 37,024 tokens and is composed of 2 sections of different styles:

- “新闻联播” *xin-wen-lian-bo* ‘news feeds’ (name of a famous Chinese TV news program), hereinafter referred to as XWLB, is a transcription of the program. The text is usually read and the style of the language is quite formal. The section contains 17,061 words.
- “实话实说” *shi-hua-shi-shuo* ‘straight talk’ (name of a famous Chinese talk show), hereinafter referred to as SHSS, is of more colloquial language type, containing spontaneous speech appearing in interviews of people of various social backgrounds, ranging from farmers to successful businessmen, The section contains 19, 963 words.

Both sections have been annotated manually as described by Liu (2006). Table 1 shows the file format of this Chinese dependency treebank, which is similar to the CoNLL dependency format, although a bit more redundant (double information on the governor’s POS) to allow for easy exploitation of the data in a spreadsheet and converting to language networks. The data can be represented as a dependency graph as shown in Figure 1.

The POS and dependency annotation is done on the transcribed texts. As the treebank contains different styles, it allows for general conclusions about the language, in spite of the limited size of the corpus. Another benefit of the double nature of the data is that we can do comparative work based on these 2 sections.

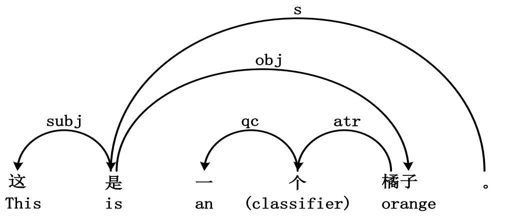

Figure 1. The graph of the dependency analysis of 这是一个橘子 *zhe-shi-yi-ge-ju-zi* ‘this is an orange’

With words as nodes, dependencies as arcs, and the frequency of the dependencies as the value of arcs, we can build a network. For example, the sample shown in Figure 1 can be converted to a network as shown in Figure 2 (excluding punctuation).

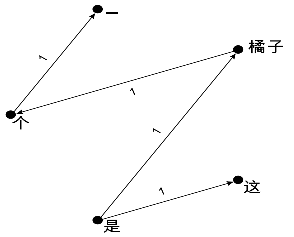

Figure 2. Network of 这是一个橘子 zhe-shi-yi-ge-ju-zi ‘this is an orange’

Following the same principle, our Chinese treebank can be presented as Figure 3, an image that gives a broad overview of the global structure of the treebank.

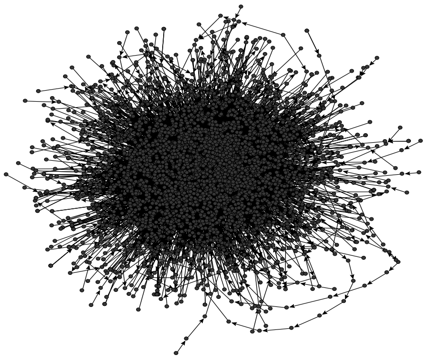

Figure 3. The network of our Chinese treebank

The resulting network has the following prop-

Sentence Order	Dependent			Governor			Dependency type
	Order	Character	POS	Order	Character	POS	
S1	1	*zhe*	pronoun	2	*shi*	verb	subject
S1	2	*shi*	verb	6	。	punctuation	main governor
S1	3	*yi*	numeral	4	*ge*	classifier	complement of classifier
S1	4	*ge*	classifier	5	*juzi*	noun	attributer
S1	5	*juzi*	noun	2	*shi*	verb	object
S1	6	。	punctuation				

Table 1. Annotation of a sample sentence in the Treebank.
这是一个橘子 *zhe-shi-yi-ge-ju-zi* ‘this is an orange’

erties: it is fully connected and there are no isolated vertices, it is a ‘small word’ and has a ‘scale-free’ structure. As we mentioned before, there are not many language characteristics that we can deduce directly from this big picture. What we need to do is to looki into the structure of some specific words in this big network, which in our study has brought about some interesting findings. The first step is to decide on the words we wanted to look into: the function words.

4 Chinese Function Words

Chinese is an isolating language: syntactic structure relies primarily on function words and word order rather than on rich morphological information to encode functional relations between elements (Levy & Manning 2003). Function words are words that have little lexical meaning or have ambiguous meaning, but instead express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker” (Klammer et al. 2000). In Chinese, function words include prepositions, conjunctions, and auxiliary and modal particles (Yu 1998).

As in any language, function words distinguish themselves not only by their syntactic properties, but also simply by their frequency. The words we are interested in are among the most common Chinese words: 在[1] *zai* ‘(to be located) in or at’, 了 *le* ‘perfective aspect marker or modal particle intensifying the preceding clause’.

We compared the frequent function words shown in XWLB, SHSS, and the *Modern Chinese Frequency Dictionary* and found that there are 3 function words that appear in all these 3 resources. They are: ‘的’ *de* ‘ablative cause suffix or possessive particle similar to the English genitive marker *’s*’, *zai* and *le*. The frequency information of these 3 function words is shown in Table 2[2].

We will exclude *de* from this study because of its unique behavior[3]. We only chose *zai* and *le* as our research objects.

The differences in distribution between the two genres of texts are mostly based on the lexical poverty of spontaneous speech (SHSS) compared to written style, resulting in higher frequencies (of the smaller number of types) in the former genre. Moreover, the notably higher relative frequency of *le* in SHSS can be explained by the fact that one usage of *le* is an intensifier typical for the genre of spontaneous oral language. Inversely, *zai* can be omitted before locatives in oral Chinese.

XWLB			SHSS			MCFD		
R	F_1	W	R	F_1	W	R	F_2	W
1	930	*de*	1	1051	*de*	1	69080	*de*
3	223	*zai*	6	429	*le*	2	26342	*le*
4	202	*le*	21	124	*zai*	6	13438	*zai*

Table 2. The frequency information of 3 function words. *R-rank, F_1-frequency, W-word, F_2-frequency in 10000,*
MCFD-Modern Chinese Frequency Dictionary[4]

5 Chinese function words in treebanks

The traditional research on Chinese function words categorizes the linguistic units, describes which words function words can connect to, and defines the relationship between function words and other linguistic units by giving some examples. This type of research has achieved valuable results and contributed to the commonly accepted classification of Chinese function words. But due to the lack of tools for collecting and processing large data, the examples are limited and most of them are not drawn from real language data either. Recently, with the appearance of corpora in Chinese linguistic research, these points improved slightly. Simple text corpora or POS tagged corpora can supply giant amounts of examples. Treebanks, however, are able to provide much richer structural information, of syntactic or semantic nature, though their size is usually rather limited. For studies on syntactic structures, as the present work on function words, treebanks are the best choice.

[1] In Chinese, *zai* may be a verb, adverb or preposition. Here we only refer to the preposition.

[2] Considering the size of XWLB and SHSS, we only paid attention on the function words whose frequency is in the top 30 of all words that have shown in these transcriptions.

[3] In Chinese, the function word ‘的’ *de* ‘’s’ is a very special word. It can pretty much follow any language unit and construct a so-called *de-structure*, *de* together with the preceding unit becoming an attribute or an expression referring to something or someone. Considering the complicated situations of the *de-structure*, they would require a special and extensive discussion, and are left for future research.

[4] In Chinese, *le* and *zai* also can be content words even though these phenomena are not common. The Modern Chinese Frequency Dictionary doesn’t distinguish these difference but we believe the deviation of the data won’t change the fact that these 2 function words are among the most common Chinese words.

This paper focuses on the structural distribution of linguistic units (words, in this study), more specifically of the function words *zai* and *le*. There is similar research on Chinese with different concerns: Liu (2007) analyzed the distribution of dependency relations and dependency distance in a Chinese treebank, including but not centered on function words. Gao (2010) and Gao et al. (2010) described the syntactic functions of nouns and verbs in mandarin Chinese, dividing syntactic functions of nouns and verbs into typical ones and atypical ones for a quantitative analysis. Chen et al. (2011) tried to build a model of valency pattern from syntactic networks based on treebanks. All these works are done from the perspective of parts of speech instead of specific words. At the same time, there are several studies in Chinese concerned with specific words. For example, Liu and Liu (2011) have engaged in a study on the evolution process of the syntactic valency of the verb. They constructed three corpora of ancient classical Chinese, ancient vernacular and the modern vernacular, and selected ten verbs as the objects of their study to ascertain the diachronic behavior of these words. Even though they analyzed the complements and modifications of the words, they failed to give specific information about the complements and modifications, only distinguishing single word units from more complicated linguistic units. In contrast, our study provided more information of the words that can connect with *zai* and *le*.

We analyzed the distribution of the dependents and governors of *zai* and *le* in XWLB and SHSS. The results are shown in Table 3 and Table 4.

Note that the genre differences are visible for both words: The governors of *zai* are very similarly diversely distributed for both genres but in spontaneous speech, the dependents of *zai* are much more diverse than in written style. Compared to *zai*, *le* has simpler combinatory possibilities (and no dependents), and here, the governors are more diverse in spontaneous speech than in written style.

Comparing these two words, we can see that, in general, *zai* can relate to more types of part of speech than *le*. However, it is not easy to interpret these tables and we will see that when passing to a network representation, the differences become much more easily accessible.

XWLB			**SHSS**		
X ——> '在' *zai*					
Gov of *zai*	Freq	%	Gov of *zai*	Freq	%
verb	208	92.86	verb	115	92.74
auxiliary	9	4.02	auxiliary	5	4.03
conjunction	2	0.89	adjective	2	1.61
adjective	1	0.45	noun	1	0.81
noun	1	0.45			
preposition	1	0.45			
pronoun	1	0.45			
'在' *zai* ——> X					
Dep of *zai*	Freq	%	Dep of *zai*	Freq	%
noun	215	96.41	noun	106	78.52
pronoun	4	1.79	pronoun	12	8.89
classifier	2	0.90	verb	8	5.93
conjunction	1	0.45	adverb	6	4.44
verb	1	0.45	auxiliary	2	1.48
			conjunction	1	0.74

Table 3. The distribution of governors and dependents of function word *zai. Freq-frequency, Dep-dependent, Gov-governor*

XWLB			**SHSS**		
X ——> '了' *le*					
Gov of *le*	Freq	%	Gov of *le*	Freq	%
verb	198	98.02	verb	384	89.51
adjective	3	1.49	adjective	38	8.86
noun	1	0.50	noun	5	1.17
			adverb	1	0.23
			classifier	1	0.23

Table 4. The distribution of governors of function word *le. Freq-frequency, Dep-dependent, Gov-governor*

6 Network properties of Chinese function words

6.1 Properties of '在' *zai* and '了' *le*

With the XWLB and SHSS syntactic networks, we studied the most frequently used network parameter of the words, the *degree*: The *degree* of a vertex (a word) refers to the number of its neighbors. This variable actually describes the number of different word types which are connected with a specific word. The directions of the arcs distinguish between *indegree* and *outdegree*. The *indegree* of a word is the number of arcs it receives while the *outdegree* is the number of arcs it sends. Reformulated linguistically, the *indegree* reflects the number of governors of a word and the *outdegree*, the number of the word's dependents. In our network, these two function words have the following properties in Table 5.

Although the size of the original sections of XWLB and SHSS in the treebank is similar (in tokens), the size of the XWLB and SHSS networks is quite different due to the difference in the lexical richness. In order to make the data more comparable, we standardized the original data, also shown in Table 5. The table clearly shows that: *le* has a zero outdegree because it cannot govern other words in our analysis of Chinese while *zai* has both indegree and outdegree; Besides, *le*'s degree is higher in SHSS than XWLB which states that the combinatory possibilities of *le* is more diverse in spontaneous speech. On the contrary the distribution of word types that *zai* can connect with is more diverse in written style, especially obvious when it comes to the indegree.

Features	'了' *le*		'在' *zai*	
	XWLB	SHSS	XWLB	SHSS
Degree	133	234	222	131
SD	0.14	0.28	0.23	0.16
Outdegree	0	0	88	61
SOD	0	0	0.17	0.12
Indegree	133	234	134	70
SID	0.29	0.55	0.29	0.16

Table 5. The degree, indegree and outdegree of the function words *zai* and *le. SD-Standard degree, SOD-Standard outdegree, SID-Standard Indegree*

6.2 Network Manipulation

To see the role that these 2 words play in the whole language network system, we carry out the following manipulations on the network: Since we are only concerned with the vertices connected to these two words, we removed all the vertices and arcs that are not connected to them. Figure 4 illustrates the graph of the remaining vertices and arcs of *zai* in XWLB. Actually, we tried to do the same thing based on the original treebank. No doubt, the idea is workable but it is difficult to visualize the result. Since the words, which are either the governors or dependents of these function words, are numerous, it would take a very big table, more than 200 lines, to show all the detailed information. A more reasonable way to visualize the data, making it more readable, is making a graphical representation of the information, such as a scatter diagram or a network diagram as the one in Figure 4.

In this diagram, we managed to arrange the words by the value of their arcs connected with *zai*. The words between circle Ⓐ and Ⓑ labeled with smallest vertices, far away from the center vertex *zai*, only connected with *zai* once in the treebank. The lines between these words and *zai* are numbered by the frequency of the connection shown in the treebank. Following the same principle, the words between circle Ⓑ and Ⓒ connected with *zai* twice in the treebank, and so they are nearer to the center vertex. The words between circle Ⓒ and Ⓓ connected *zai* three times and the words in circle Ⓓ, except the word *zai* itself, connected with *zai* more than three times in the treebank. The more connections there are between the words and *zai*, the bigger the size of the vertices representing the words, the shorter the distance between the words and *zai*. In this way, the diagram 4 clearly shows that, even though *zai* has many neighbors, most of them seem to prefer visiting it just once or twice, in other

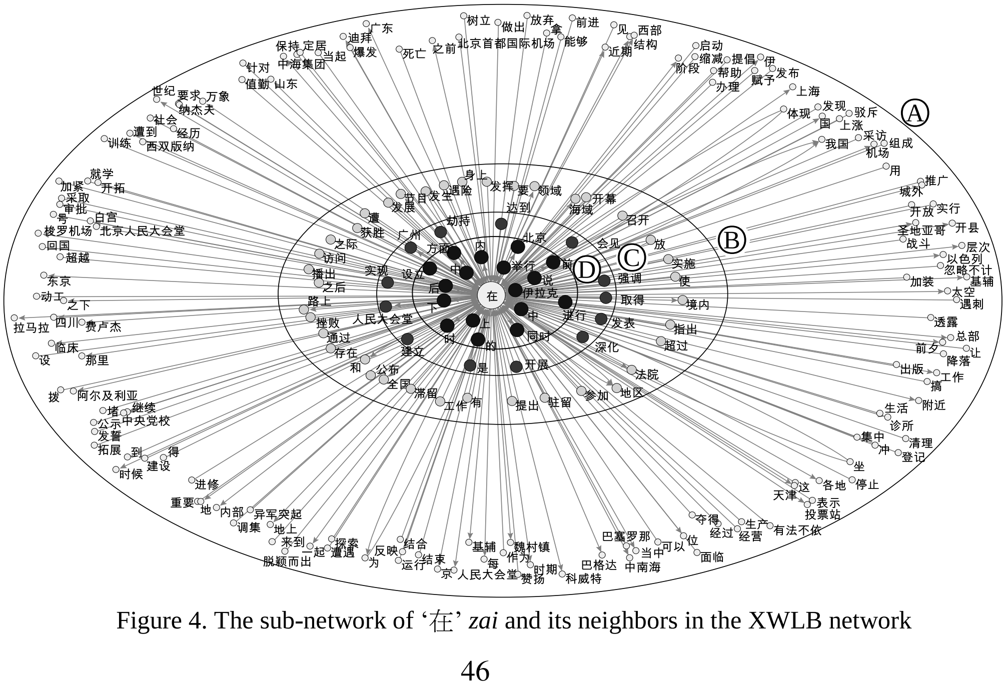

Figure 4. The sub-network of '在' *zai* and its neighbors in the XWLB network

words, the number of connection is distributed more evenly among its neighbors.

After removing the vertices, we combined all the words with the same part of speech except the two words we are studying. So we got a new “mixed language network”. It mixes two types of vertices, one representing a word while the other one representing the part of speech. This new graph, as shown in Figure 5, also included the information of Table 3 and Table 4. The results we got from analyzing the treebank can also be extracted from the language networks.

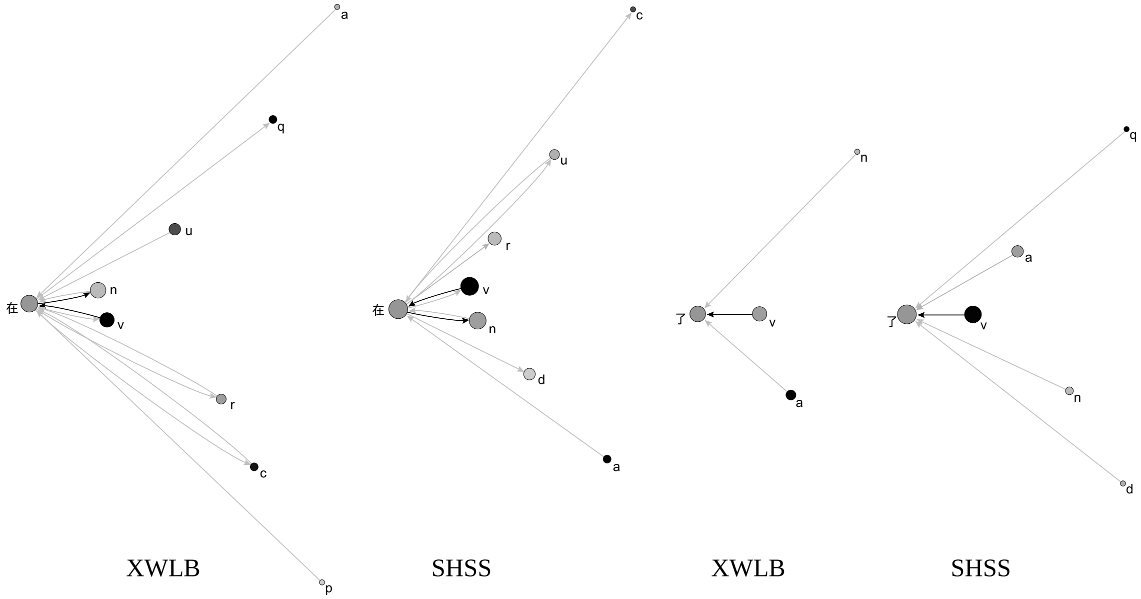

Figure 5. The distribution of governors and dependents of the function words *zai* and *le*

In this diagram, we put the arcs in different grayscales. The higher the value of an arc is, i.e. the frequency in Table 3 and Table 4, the darker their color, the bigger size of the vertices which represent the parts of speech. It is even easier to get the same conclusion than that drawn from Table 3 and Table 4: *le* can only be a dependent while *zai* can be a governor and a dependent; *zai* can relate to more types of part of speech than the *le*, *zai* can related to more words in XWLB than in SHSS while *le* can related to more words in SHSS than in XWLB.

Now, we can see that the main difference between analysing the original treebank and the language network is that the language network can provide an easier and direct access to a graphic output, especially when the data is too complex or too big to be included in a limited table.

So are there other facts about the language structure that are extremely difficult to see directly from the original treebank and that can be seen directly in the network? One of the advantages of the language network model is that it views the language as a connected whole system. Without the language network approach, describing the language system is more like talking about an unspecified abstract structure. The language network model gives a more specific structure model to the language system and also provides different computational tools that have proven to be successful in sociology and computer science, which are able to describe the different elements of a network system, or, as in our case, a language system. So we tried to manipulate the XWLB and SHSS networks to find out the roles of these two function words in the language networks systems. The way we tried actually follows a very simple logic. If you want to know the function of one element in a system, the simplest way is to remove it from the system and then to see what the consequences are: We respectively removed the vertices representing *zai* and *le* from XWLB and SHSS language networks and compared several most common features of the networks, *the number of vertices*, *average degree*, *the number of isolated vertices*, before and after removing the vertex.

The numbers of vertices are actually the numbers of word types in the treebank. Although the sizes of XWLB and SHSS are similar, the numbers of vertices of XWLB and SHSS networks, or the size of the networks, are obviously different due to the difference of lexical richness.

Network		Num	IV	AD
XWLB	Original	4011	0	6.15
	le Removed	4010	0	6.09
	zai Removed	4010	17	6.04
SHSS	Original	2601	0	8.56
	le Removed	2600	0	8.38
	zai Removed	2600	5	8.46

Table 6. The network data before and after removing the function words. *Num: Numbers of vertices, IV: Isolated Vertices, AD: Average Degree*

The isolated vertices represent the vertices without any neighbors. This is the interesting part here. According to the data, there are no isolated vertices after removing *le*. All the remained vertices are still fully connected. So, if we believe the network somehow can be seen as the model of the syntactic structure of the language system drawn from this part of the treebank, then removing *le* seems to cause no significant trouble here. The whole structure didn't suffer from a systematic crisis, even though the *le* was a high frequency word with very high degrees. At the same time, removing *zai* caused isolated vertices in both XWLB and SHSS networks, especially in SHSS, even though the *zai* has lower frequency than *le* in the treebank and lower degrees in the network. In other words, removing this word created a much bigger systematic crisis. The reason is simple: *le* can only be a dependent. Take a picture like diagram 6: In the simple full connected network there is a vertex A that only has indegree and no outdegree. Because vertex A only attaches to other vertices and it doesn't convey any unique information between its neighbors, removing it from the network won't render any vertex isolated.

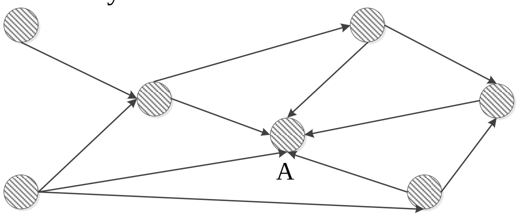

Figure 6. A simple network example

This result fits a common sense in syntax that the governors are somehow more important than dependents when it comes to the structural completion of sentences. But it is very difficult to quantify the syntactic importance, especially for the whole treebank, text or language systems. We see that the analyzed function words, which share high frequency and degrees, in fact play very different roles in the system model: As a result, it seems safe to claim that *zai* is more important than *le* for this model's structure. The syntactic importance of specific words can be quantified in this way. Developing a numeric scale of a well-defined notion of "syntactic importance" is left for future research.

This study shows that the language network approach can not only provide an easier and direct access to getting a graphic output but also can bring some fresh new angles for language analyzing.

7 Conclusion

This paper addresses the importance of developing techniques of treebank exploitation for syntactic research ranging from theorem verification to discovery of new relations invisible to the eye.

We advocate in particular the usage of network tools in this process and show how a treebank can, and, in our view, should be seen as a unique network.

We have shown in more detail, by opposing the function words *zai* and *le*, that the frequency of words is not equivalent to the word's importance in the syntactic structure, pointing to a notion that we may call the "centrality" of the word. The importance in the syntactic structure is still a vague notion that needs to be refined further, but simple network manipulations like removal of the words in question can reveal properties of the words that seem to be closely related to the words' structural roles. For example, a word A whose removal breaks the network in parts is clearly more important than a word B whose removal preserves the connectedness of the network (as the word only occupies exterior nodes). Since the results shown in this paper confirm well-known facts concerning these two function words, the same method can be applied to other function words as well content words. Ongoing research includes analyses of the Chinese equivalent of the following words: *de* 'ablative cause suffix or possessive particle similar to the English genitive marker 's', *wo* 'I, me, myself', *shi* 'are, am, yes', *ge* 'individual, entries', *yi* 'one, single', *zhe* 'this, it,these', *bu* 'do not, need not', *ta* 'he, him', *shuo* 'speak, talk, say', *ren* 'person, people, human being', and *dao* 'arrive, reach, get to'.

We leave it for further research to develop the notion of "centrality" into a numerical value that would allow comparing any pair of words.

Equally, the active field of network analysis will in time reveal new techniques that have in turn to be applied to new and bigger language networks based on treebanks of different types and languages. This could establish network syntax as one branch of the emerging field of data-driven linguistics.

References

Abramov O. and Mehler A. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4), 291-336.

Allee V. 2000. Knowledge networks and communities of practice. *OD Practitioner Online*, 32(4), 1-15.

Amancio D. R., Antiqueira L., Pardo T. A. S., da F. Costa L., Oliveira Jr O. N. and Nunes M. G. V. 2008. Complex networks analysis of manual and machine translations. *International Journal of Modern Physics C*, 19(04), 583-598.

Amancio D. R., Nunes M. G. V., Oliveira Jr O. N., Pardo T. A. S., Antiqueira L. and da F Costa L. 2011. Using metrics from complex networks to evaluate machine translation. *Physica A*, 390(1), 131-142.

Arbesman S., Strogatz S. H. and Vitevitch M. S. 2010. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos*, 20(03), 679-685.

Baker M. 1993. Corpus linguistics and translation studies: Implications and applications. *Text and technology: in honour of John Sinclair*, 233, 250.

Banisch S., Araújo T. and Louçã J. 2010. Opinion dynamics and communication networks. *Advances in Complex Systems*, 13(01), 95-111.

Barabási A. L. and Bonabeau E. 2003. Scale-free networks. *Scientific American*, 288(5), 50-9.

Böhmová A., Hajič J., Hajičová E., et al. 2003. *The Prague dependency treebank*. In *Treebanks*, 103-127. Springer, Netherlands.

Borgatti S. P., Everett M. G. and Freeman L. C. 2002. *Ucinet for Windows: Software for social network analysis*. Analytic Technologies, Harvard.

Borgatti S. P. 2002. *NetDraw: Graph visualization software*. Analytic Technologies, Harvard.

Borge-Holthoefer J. and Arenas A. 2010. Semantic Networks: Structure and Dynamics. *Entropy*, 12(5), 1264-1302.

Čech R. and Mačutek J. 2009. Word form and lemma syntactic dependency networks in Czech: A comparative study. *Glottometrics*, 19, 85-98.

Čech R., Mačutek J. and Žabokrtský Z. 2011. The role of syntax in complex networks: Local and global importance of verbs in a syntactic dependency network. *Physica A*, 390(20), 3614-3623.

Charteris-Black J. 2004. *Corpus approaches to critical metaphor analysis*. Palgrave-MacMillan.

Chen X. and Liu H. 2011. Central nodes of the Chinese syntactic networks. *Chinese Science Bulletin*, 56(1): 735-740.

Chen X., Xu C. and Li W. 2011. Extracting Valency Patterns of Word Classes from Syntactic Complex Networks. *Proceedings of Depling 2011, International Conference on Dependency Linguistics*. Barcelona, 165-172.

Christiano Silva T. and Raphael Amancio D. 2013. Network-based stochastic competitive learning approach to disambiguation in collaborative networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 23(1), 013139-013139.

Dickinson M. and Meurers W. D. 2003. Detecting inconsistencies in treebanks. *Proceedings of TLT*, 3, 45-56.

Ferrer i Cancho R. 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. *Problems of quantitative linguistics*, 60-75.

Ferrer i Cancho R., Capocci A. and Caldarelli G. 2007. Spectral methods cluster words of the same class in a syntactic dependency network. *International Journal of Bifurcation and Chaos*, 17(07), 2453-2463.

Ferrer i Cancho R., Solé R. V. and Köhler R. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69(5), 051915.

Gao S. 2010. A Quantitative Study on Syntactic Functions of Nouns in Mandarin Chinese: Based on Chinese Dependency Treebank. *TCSOL Studies*, 2, 54-60.

Gao S., Yan W. & Liu H. 2010. A Quantitative Study on Syntactic Functions of Chinese Verbs Based on Dependency Treebank. *Chinese Language Learning*, 5, 105-112.

Gong T., Baronchelli A., Puglisi A. and Loreto V. 2012. Exploring the role of complex networks in linguistic categorization. Artificial Life, 18(1), 107.

Ke J. and Yao Y. A. O. 2008. Analysing Language Development from a Network Approach. *Journal of Quantitative Linguistics*, 15(1), 70-99.

Klammer T. P., Klammer T. P., Schulz M. R. and Della Volpe A. 2000. *Analyzing English Grammar*, 6^{ed}. Pearson Education India.

Kretzschmar W. A. 2009. The linguistics of speech. Cambridge University Press, New York.

Levy R. and Manning C. 2003. Is it harder to parse Chinese, or the Chinese Treebank?. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 1, 439-446. Association for Computational Linguistics.

Li J. and Zhou J. 2007. Chinese character structure analysis based on complex networks. *Physica A*, 380, 629-638.

Li Y., Wei L., Li W., Niu, Y. and Luo S. 2005. Small-world patterns in Chinese phrase networks. *Chinese Science Bulletin,* 50(3), 287-289.

Liu B. & Liu H. 2011. A Study on the Evolution of the Verbal Syntactic Valence Based on Corpus. *Language Teaching and Linguistic Studies*, 6, 83-89.

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

Liu H. 2007. Probability distribution of dependency distance. *Glottometrics*, 15, 1-12.

Liu H. 2008. The complexity of Chinese dependency syntactic networks. *Physica A*, 387, 3048-3058.

Liu H. 2009a. Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics*, 16(3), 256-273.

Liu H. 2009b. Statistical Properties of Chinese Semantic Networks. *Chinese Science Bulletin*, 54(16), 2781-2785.

Liu H. 2010. Language Clusters based on Linguistic Complex Networks. *Chinese Science Bulletin*, 55(30), 3458-3465.

Liu H. and Cong J. 2013. Language clustering with word cooccurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10), 1139-1144.

Liu H., Hudson R., and Feng Z. 2009. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*, 5(2), 161-174.

Liu H. and Xu C. 2011. Can syntactic networks indicate morphological complexity of a language?. *EPL (Europhysics Letters)*, 93(2), 28005.

Liu Z. Y. and Sun M. S. 2007. Chinese word co-occurrence network: its small world effect and scale-free property. *Journal of Chinese Information Processing*, 21(6), 52-58.

Lopez A., Nossal M., Hwa R. and Resnik, P. 2002. *Word-level alignment for multilingual resource acquisition*. Maryland Univ College Park Inst For Advanced Computer Studies.

Marcus M. P., Marcinkiewicz M. A. and Santorini B. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), 313-330.

Mehler A., Diewald N., Waltinger U., Gleim R., Esch D., Job B., ... and Blanchard, P. 2011. Evolution of Romance language in written communication: Network analysis of late Latin and early Romance corpora. *Leonardo*, 44(3), 244-245.

Mehler A., Lücking A. and Menke P. 2012. Assessing cognitive alignment in different types of dialog by means of a network model. *Neural Networks*, 32, 159-164.

Mehler A., Lücking A. and Weiß P. 2010. A network model of interpersonal alignment in dialog. *Entropy*, 12(6), 1440-1483.

Mille S. and Wanner L. 2010. Syntactic dependencies for multilingual and multilevel corpus annotation. *Proceedings of LREC 2010*. Malta.

Mukherjee A., Choudhury M., Ganguly N. and Basu A. 2013. Language Dynamics in the Framework of Complex Networks: A Case Study on Self-organization of the Consonant Inventories. In *Cognitive Aspects of Computational Language Acquisition*, Springer, Netherlands, 51-78.

Nooy W., Mrvar A. and Batagelj V. 2005. *Exploratory Network Analysis with Pajek*. Cambridge University Press, New York.

Peng G., Minett J. W. and Wang W. S. Y. 2008. The networks of syllables and characters in Chinese. *Journal of Quantitative Linguistics*, 15(3), 243-255.

Shannon P., Markiel A., Ozier O., Baliga N. S., Wang J. T., Ramage D., ... and Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), 2498-2504.

Watts D. J. and Strogatz S. H. 1998. Collective dynamics of ‘small-world’ networks. *nature*, 393(6684), 440-442.

Yu S. 1998. *Modern Chinese grammatical information dictionary explanation*. Tsinghua university, Beijing.

Yu S., Liu H. and Xu C. 2011. Statistical properties of Chinese phonemic networks. *Physica A*, 390(7), 1370-1380.

Zeldes A., Ritz J., Lüdeling A. and Chiarcos C. 2009. ANNIS: A search tool for multi-layer annotated corpora. *Proceedings of corpus linguistics*, 9.