Poznań Studies in Contemporary Linguistics 45(4), 2009, pp. 495–509

doi:10.2478/v10010-009-0025-3

CHINESE SYNTACTIC AND TYPOLOGICAL PROPERTIES BASED ON DEPENDENCY SYNTACTIC TREEBANKS

HAITAO LIU[1], YIYI ZHAO AND WENWEN LI
Communication University of China
[1]lhtcuc@gmail.com

ABSTRACT

This paper offers a quantitative analysis of the syntactic and typological properties of Chinese based on five Chinese dependency treebanks. The study shows that mean dependency distance of Chinese is 2.84; 40–50% dependencies are between non-adjacent words; Chinese is a mixed language with a governor-final and SV–VO–AdjN preference; the mean dependency distance of governor-initial dependencies is greater than that of governor-final ones. Methodologically, the paper adopts five treebanks with different text genres and annotation schemes as a resource to study syntactic features of a language. This method avoids corpus influences on results so that the conclusions can be more reliable and robust. If suitable treebanks are available, it will be an easy task to apply our method to other languages. In this way, the method has a broad theoretical and cross-linguistic perspective.

KEYWORDS: Chinese; dependency distance; dependency direction; dependency treebank; linguistic typology.

1. Introduction

As a subject which adopts exact and empirical methods to study human languages, quantitative (corpus) linguistics has progressed considerably in recent years (Köhler et al. 2005; Best 2006; Bod et al. 2003; Gries 2009). At present, besides the statistic of traditional word frequency, quantitative and corpus linguists have entered into other fields of linguistic research, such as syntax, pragmatics, typology and language evolution. For these quantitative analyses, some analyzed or annotated language corpora are needed. However, it is very time-consuming to build large amounts of corpora that contain syntactic and other linguistic knowledge. Therefore, we should turn our view to other linguistic fields which have similar needs to look for the possibility of sharing linguistic resources.

Numerous corpora with syntactic annotation have been built since the prevailing of corpus-based method in computational linguistics. A corpus of this kind is commonly called a "treebank" (Abeillé 2003). In 1993, Pennsylvania University built the first English phrase structure treebank (Marcus et al. 1993). Later, treebanks of different languages came out one after another and the tendency of construction of treebanks has been changed. For example, the annotation scheme turns gradually from phrase structure to dependency structure (Kakkonen 2005); and treebank annotation combines closer with linguistic theories (De Smedt et al. 2007). These changes make treebanks no longer just as a tool for training and evaluating a syntactic parser in computational linguistics, but resources for quantitative and empirical language research.

Köhler and Altmann (2000) quantitatively analyzed English syntax features based on a phrase structure treebank (the Susanne corpus), and built the fundamentals of quantitative study of syntax. The greatest difference between the quantitative study of syntax and that of word frequency, word length, sentence length, etc., is that syntactic research closely relates to syntactic theory. In order to better discover the properties of syntactic structure of human language, it is not enough to only use syntactic theories based on phrase structure. Moreover, the structure of a syntactic treebank is influenced by text genre and annotation schemes, so that it is very difficult to draw a convincing conclusion based on only one treebank even when we adopt the same syntactic model to do a quantitative study of only one language. In order to more accurately find those quantitative syntactic characteristics of natural language, we need to adopt several treebanks to study one linguistic phenomenon of a certain language.

Dependency analysis is another kind of competitive syntactic analysis method apart from phrase structure (or constituent analysis). Thereby it is necessary to quantitatively explore syntactic characteristics of language based on dependency treebanks. Such studies are helpful to discover several linguistic features which phrase structure treebanks can not discover. Liu (2009a) investigated probability distributions of dependency relations extracted from a Chinese dependency treebank and found that the most investigated distributions are excellently fitted with modified right-truncated Zipf-Alekseev distribution. This study is a very useful beginning for quantitative research based on dependency treebanks. Using five Chinese dependency treebanks, the paper quantitatively examines several linguistic features of Chinese, such as dependency distance and dependency direction.

Section 2 introduces the basic concepts and statistical methods of dependency grammar, dependency distance and dependency direction. Section 3 briefly introduces the five dependency treebanks used in this study and gives the statistical results for the relevant quantitative items. Section 4 is an analysis and discussion of the results. The last section summarizes our findings and puts forward the directions of further work.

2. Dependency grammar, dependency distance and dependency direction

Linguists still differ in their understanding of what a dependency-based grammar is. However, the following properties, which are generally accepted by linguists, are considered as the core features of a syntactic dependency relation (Tesnière 1959; Mel'čuk 1988; Hudson 2007; Liu 2009b):

- It is a binary relation between two linguistic units.
- It is usually asymmetrical and directed, with one of the two units acting as the governor and the other as the dependent.
- It is labeled, and the type of a dependency relation is usually indicated using a label on top of the arc linking the two units.

These features can be diagrammatically shown as Figure 1.

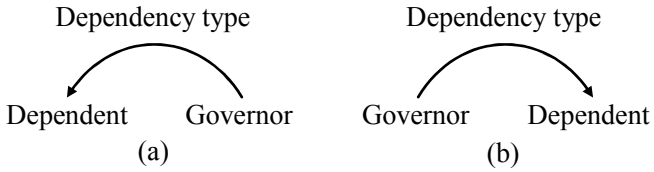

Figure 1. Three elements of a dependency.

Figure 1 shows a dependency relation between Dependent and Governor, whose label is "Dependency type". The directed arc from Governor to Dependent demonstrates the asymmetrical relation between the two units. (a) and (b), respectively, denotes governor-final and governor-initial, which reflects the linear order of the two components in the sentence. We call (a) governor-final dependencies, and (b) governor-initial dependencies. Dependency analysis can be seen as a set of all dependency relations existing in a sentence.

Example (1) is a simple sentence which can be used to illustrate the analysis.

(1)

这	是	一	个	例子
Zhe	shi	yi	ge	lizi
This	is	one	classifier	example

'This is an example.'

Figure 2 (overleaf) is the dependency analysis of example (1).

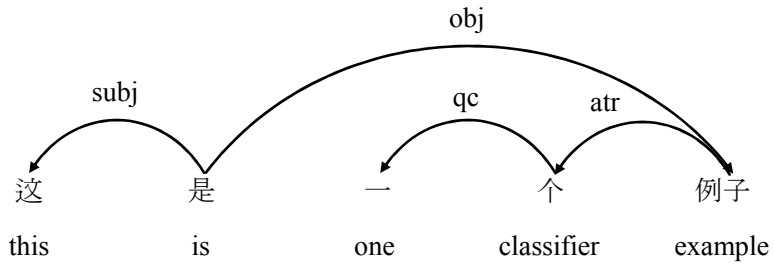

Figure 2. The dependency analysis of a Chinese sentence as a graph.

In order to better use the syntax information in Figure 2, we can convert it into another format which is convenient for computer operation, as in Table 1.

Table 1. Annotation of dependency analysis in table format.

Dependent			Governor			Dependency type
Order	Word	POS	Order	Word	POS	
1	这	r	2	是	v	subj
2	是	v				
3	一	m	4	个	q	qc
4	个	q	5	例子	n	atr
5	例子	n	2	是	v	obj

Every row in Table 1 corresponds to a dependency relation in dependency analysis structure. A sentence which contains n words has $n-1$ dependency relations. If we analyze m sentences in a language with this method, then it will form $(m \times n) - m$ dependency relations, the collection of which produces a dependency treebank.

This paper mainly studies the dependency distance and dependency direction of Chinese. Dependency distance is the linear distance between governor and dependent. The term "dependency distance" was introduced by Hudson (1995: 16), who defined dependency distance as "the distance between words and their parents, measured in terms of intervening words". That is to say, dependency distance between two adjacent words is zero; and that with an interval word is one.

Although Hudson's definition shows the essential attributes of dependency distance, it will be a little inconvenient when used in large-scale automatic statistics. Therefore, we define dependency distance as the difference that governor order deducts from dependent order. The governor may occur before or after the dependent; therefore, the difference could be negative or positive, which is marked as the direction of a dependency relation, or "dependency direction" for short. We call the nega-

tive dependency relations "governor-initial dependencies", and the positive relations as "governor-final dependencies". For example, in the sentence in Figure 2 and Table 1, the dependency distance of "是–这" is 2−1=1, "个–一" is 4−3=1, "例子–个" is 5−4=1, all of which are positive, while the dependency distance of "是–例子" is 2−5=−3, which is negative. That is to say, this sentence contains three governor-final dependency relations and one governor-initial.

We can also use the absolute value of dependency distance to calculate mean dependency distance (hereafter, MDD) of a sentence or a treebank. MDD of the sentence in (1) is defined in (2):

$$\text{(2)}\quad \text{MDD(the sentence)}=\frac{1}{n-1}\sum_{i=1}^{n-1}|DD_i|$$

Here, *n* is the number of words in the sentence. *DDi* is the dependency distance of the *i*-th syntactic link of the sentence. According to this formula, the MDD of the above example is: (1+1+1+3)/4=1.5.

Formula (2) can also be extended to calculate the mean dependency distance of a larger collection of sentences, such as a treebank:

$$\text{(3)}\quad \text{MDD(the sample)}=\frac{1}{m-n}\sum_{i=1}^{m-n}|DD_i|$$

In formula (3), *m* is the total number of words in the sample, *n* is the total number of sentences in the sample. *DDi* is the dependency distance of the *i*-th syntactic link of the sample. Liu et al. (2009) provided a detailed method to measure dependency distance based on a treebank.

3. Statistics of Chinese dependency distance and dependency direction

A treebank is a corpus with syntactic annotation. Quantitative linguistic study based on a treebank may be influenced by the schemes of syntactic annotation, size of the treebank, text genre of corpus, etc. In order to more accurately grasp the quantitative syntactic features of a language, it is not enough to only use one treebank.

The present paper chooses the Sinica Treebank (hereafter *sinica*, Chen et al. 2003), the Penn Chinese Treebank (hereafter *pct*, Xue et al. 2005), the Chinese dependency treebank by IR-lab at HIT (hereafter *hit*, Ma 2007), and two dependency treebanks built by the authors' group (hereafter *cucc* and *cucn*, Liu 2009b). Among these five treebanks, *hit*, *cucc* and *cucn* are dependency treebanks; *sinica* uses the dependency treebank format of Shared Task on Multilingual Dependency Parsing

(Kübler et al. 2009); and *pct* is a dependency format provided by Joakim Nivre. In order to use the methods to calculate the dependency distances and dependency directions mentioned above, we converted all five treebanks into the format in Table 1. By using formula (3), we get the results in Table 2.

Table 2. Statistic results for five Chinese dependency treebanks.
(Raw figures are included in parentheses.)

Treebank	*sinica*	*cucc*	*hit*	*cucn*	*pct*
Size	280205	19060	168470	16654	412191
msl	4.9	20.6	17	24	22.9
Pdd%	70.2 (196790)	59.4 (11317)	66.6 (112123)	68.5 (11411)	75 (309291)
Ndd%	29.8 (83415)	40.6 (7743)	33.4 (56347)	31.5 (5243)	25 (102900)
1dd%	56.6 (158574)	53.7 (10234)	50.8 (85568)	56.3 (9378)	47.9 (197290)
2dd%	43.4 (121631)	46.3 (8826)	49.2 (82902)	43.7 (7276)	52.1 (214901)
Pmdd	1.849	1.94	2.34	2.25	3.385
Nmdd	2.423	3.96	3.99	4.3	4.49
MDD	2.02	2.76	2.89	2.89	3.66
95%	5	9	9	10	12
Genre	mixed	dialogue	news	news	news
Type	CF	D	D	D	C

Size refers to the number of dependency relations in the treebank; *msl* is the mean sentence length; *Pdd%* is the percentage of governor-final dependency relations; *Ndd%* is the percentage of governor-initial dependency relations; *1dd%* is the percentage of dependency relations in adjacent words; *2dd%* is the percentage of dependency relations in non-adjacent words; *Pmdd* is the mean dependency distance of all governor-final dependency relations; *Nmdd* is the mean dependency distance of all governor-initial dependency relations; *MDD* is the mean dependency distance of the treebank; *95%* refers to the mean dependency distance when the number of dependency relations reaches 95%; *Genre* is the type of the corpus in the treebank; *Type* shows the native annotation scheme of the treebank, where D is dependency structure, and C is constituent structure. CF is a mixed structure of constituent structure and grammatical functions.

We will discuss the data in Table 2 and the relationships among these elements in the next section.

4. Discussion

The study of dependency distance and dependency direction is beneficial for:

(1) Predicting syntactic difficulty (Gibson 1998; Liu 2008);
(2) Recognizing the mechanisms of child language learning (Ninio 2006);
(3) Exploring language typologically (Tesnière 1959; Liu in press);
(4) Designing better parsing algorithms for natural language processing (Collins 1996; Buch-Kromann 2006).

If we regard dependency distance as a criterion for measuring comprehension difficulty of a sentence, for any human language, its MDD should be in a certain range, which will be influenced by the text genre, schemes of syntactical annotation, length of sentence, etc. In general, it will be restricted by human capacity of working memory (Liu 2008).

Liu (2007a) showed that the probability distribution of dependency distance can be well captured by the right-truncated Zeta distribution, but the result is only based on six texts extracted from a Chinese dependency treebank. The distribution of dependency distance of a treebank seems more complicated and mixed. Figure 3 shows the empirical distributions of dependency distance in five treebanks.

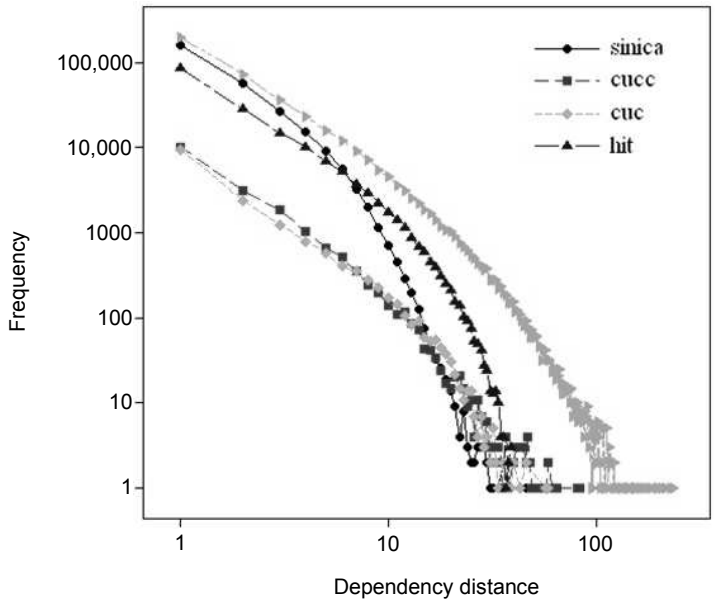

Figure 3. Empirical distributions of dependency distance in five treebanks (log x log).

One of the aims of adopting five different treebanks to study Chinese is to further investigate the factors which affect the MDD of a language.

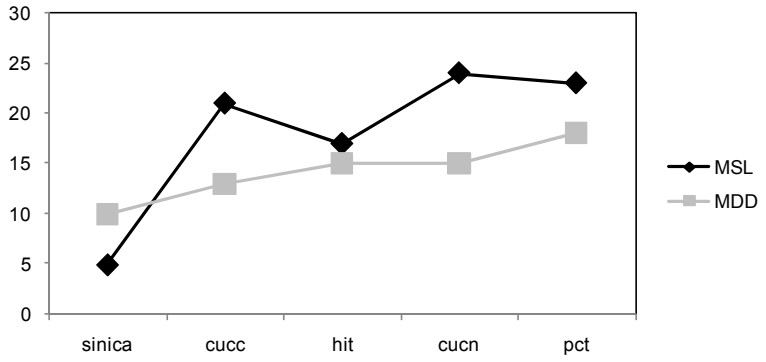

Figure 4. Relationship between MDD and mean sentence length (MSL).
MDD was multiplied by 5 for a clearer trend.

Figure 4 shows that the MDDs of these five treebanks are different, ranging from 2.02 to 3.66. In other words, the governor and dependent, which form a dependency relation in Chinese, are at a distance of 1 to 2.5 words from each other. The relation between sentence length and dependency distance can be concluded from this figure. *Sinica* has the minimum sentence length as well as the shortest dependency distance, while the dependency distances of the other four treebanks show that sentence length is merely a main factor affecting dependency distance, but not the only one. We can also see that the MDDs of these five treebanks change in a small range when they are affected by numerous factors; however, the value does not go beyond human capacity of working memory (7±2) (Miller 1956) or 4 (Cowan 2005), which indicates that dependency distance of human language has a minimizing tendency (Liu 2007a, 2008).

In order to clearly discover the relation between MDD and the other factors in Table 2, we analyze the correlative degree among these factors, as in Figure 5.

Figure 5 shows that the factors with high positive Pearson correlation coefficients with dependency distance are *95%*, *pmdd*, *nmdd*, *2dd* and *msl*, while a negative correlation is seen for *1dd%*. As the relations between *pmdd*, *nmdd*, *95%* and dependency distance are easy to understand, we will not discuss them further here. It is noticeable that there is no close relation between the size of a treebank and dependency distance, which denotes that we could use treebanks with different sizes to statistically study the features of dependency syntax. Although Figure 5 provides some visual presentation of the relationship among dependency and other factors, statistically, they do

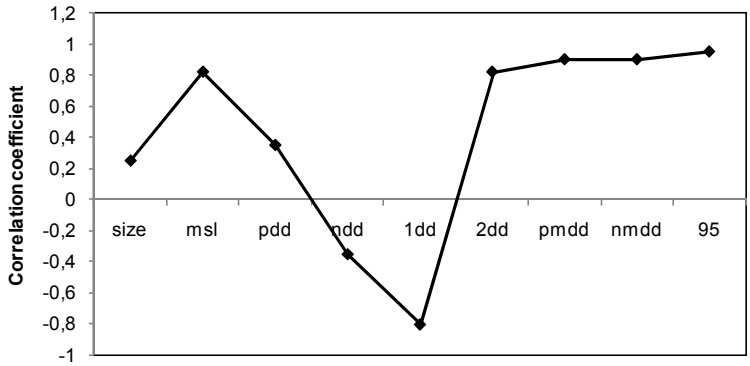

Figure 5. The correlativity among dependency distance and other factors.

not have the same significance and P-values. The statistical test shows that their P-values are as follows: *95%* (0.005), *nmdd* (0.037), *pmdd* (0.036), *1dd%* and *2dd%* (0.085), *msl* (0.088), *pdd%* and *ndd%* (0.569), *size* (0.644).

Among these five treebanks, the dependency distance of *pct* is the greatest. It is not possible to explain this by reference to mean sentence length and text genre, as *cucc* and *cucn* have similar mean sentence length with *pct*, and *cucn* and *hit* have similar text genre with it. The one and only reason for the high dependency distance in *pct* is that *pct* is a dependency treebank which is automatically converted from a phrase structure treebank. It indicates that the annotation scheme of a treebank has some effects on MDD. The MDDs of *cucc*, *cucn* and *hit* are close to each other since all of the treebanks are built according to dependency syntax, though the genre of *cucc* is conversational, *cucn* is TV news, and *hit* is newspaper. If the judgment is reasonable, a dependency treebank which is automatically converted from a phrase structure treebank may not be suitable as a resource for linguistic study. To make the treebank available for linguistic research, we have to manually correct it. This conclusion is compatible with Liu (2007b), which also considered that dependency treebanks of such a kind are not fit for linguistic study. The other four treebanks have similar MSLs, except for *sinica*. The reason *sinica* has a shorter mean sentence length is that it wipes off all punctuation in the original corpus, and breaks off the sentence whenever the punctuation appears, which makes annotating easier but loses the information in the original text and does not reflect the truth of the Chinese sentence. Therefore, we could ascribe *sinica*'s very small MDD to the special annotating scheme. In this way, if we ignore *pct* and *sinica* when we analyze dependency distance, we will get a more accurate Chinese MDD, which is about 2.84, approximately corresponding to the mean MDD of the five treebanks.

The percentage of adjacent dependency relations (1dd%) is closely related to dependency distance. The adjacent dependency relation is formed between two ad-

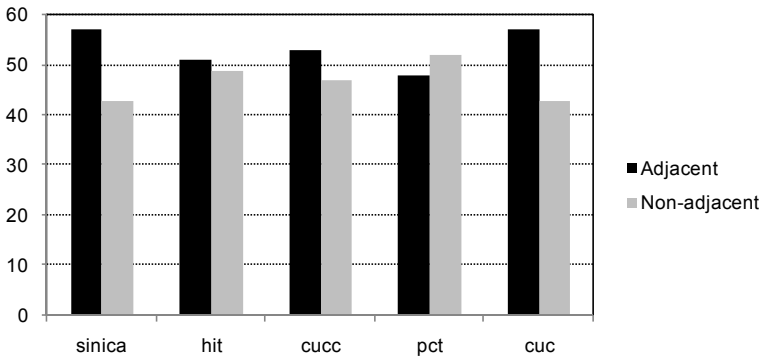

Figure 6. Percentages of adjacent and non-adjacent dependency relations.

jacent words in a sentence and its dependency distance is 1. Obviously, the more adjacent dependency relations a language has, the smaller its MDD will be. They are in a negative correlation.

Figure 6 shows that 40–50% dependencies are between non-adjacent words in Chinese. We have not discovered the main reasons through correlation analysis. Maybe we need to look for the answer through the universal of human cognitive structure. What we have found is that the proportion of dependency relations between adjacent words is approximately the same when a treebank uses different text genres and annotation schemes. The statistical tests show that the differences among percentages of adjacent dependency relations in the five treebanks are not significant ($p>0.39$).

The method for measuring dependency distance that we put forward in this paper can help us recognize whether a language is governor-final, governor-initial, or mixed. In linguistics, this approach has roots in Tesnière (1959: 22–23, 32–33). The method of using the linear order of two units constructing a grammar function as index of language typology often appears in the works of modern typology (Greenberg 1963; Dryer 1992).

What is different from the study of typology in our method based on treebanks is that it can comprehensively investigate all the linear orders of grammatical functions. And the conclusion we get is based on real language usage, not only on some representative sentences. Therefore, it can better reflect the truth of language structure. Liu (in press) statistically analyzed dependency direction of 20 languages, which proves that the quantity of dependency direction based on a treebank can be regarded as a means to study language classification and typology. If dependency direction can be a means of language classification, the proportions of governor-final and governor-initial dependency relations will not be overly high when different treebanks are adopted for a language. Figure 7 proves our view.

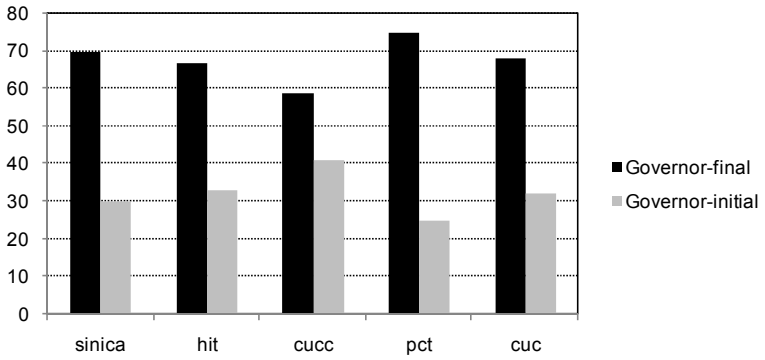

Figure 7. Percentages of dependency directions in Chinese.

Figure 7 shows that Chinese is a mixed language with a governor-final preference; all five treebanks display this tendency with few differences. Compared with Liu (in press), Chinese, Japanese, Turkish, and Hungarian display a governor-final preference. Compared with the other three languages, Chinese is weaker in this tendency, while compared with some typical balanced (mixed) languages, such as English, Dutch, or Slovenian, Chinese is stronger. *Cucc*, which uses the conversation genre, displays some differences to the other written language treebanks. What still seems unexplainable is whether these differences can be seen as a criterion to judge text genre, though the syntactic feature is the most obvious difference between conversation and news treebanks. This problem is worthy of further study. The statistical tests show that the differences among the percentages of dependency directions in the five treebanks are not significant ($p>0.17$).

Dryer (1997) considered that binary relations built on SV vs. VS, OV vs. VO are a more effective typological study feature. After demonstrating that a treebank which used to be a resource of computational linguistics can also be a resource of linguistic (typological) research, we chose these four treebanks (without *sinica*) for further study. The reason we exclude *sinica* is that we could hardly pick up the typological features we need from it. We selected the grammatical relations between *subject*, *object*, *adjective* and *noun*. Based on the definition and dependency direction calculation method mentioned above, we can easily get Table 3 (overleaf) from the treebanks.

Table 3 shows that Chinese is a language with a SV–VO–AdjN preference, which is consistent with conclusions reached by topologists using other databases (Haspelmath et al. 2005).

The other issue we are interested in is the difference between the MDDs of governor-final and governor-initial dependencies. The reason we study this is that, semantically, governor is always regarded as head word, while dependent as the complement component.

Table 3. Percentages of some specific typological features.

	VS	SV	VO	OV	Nadj	AdjN
cucc	0.7	99.3	99.6	0.6	0	100
hit	0	100	99.9	0.1	0	100
cucn	1.3	98.7	98	2	0.4	99.6
pct	0.1	99.9	100	0	0	100

VS is the percentage of verb before subject, SV is the percentage of verb after subject, VO is the percentage of object after verb, OV is the percentage of object before verb, NAdj is the percentage of adjective after noun, and AdjN is the percentage of adjective before noun.

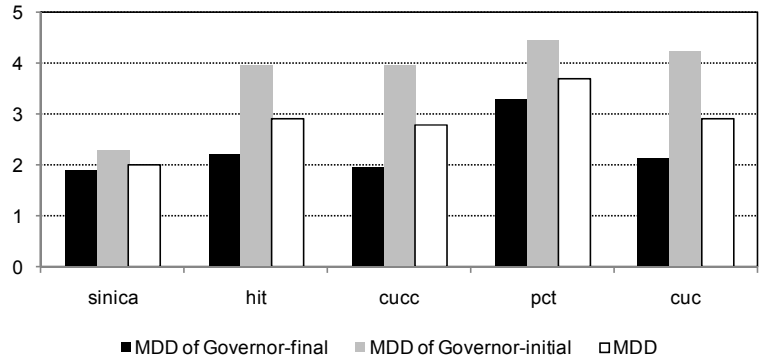

Figure 8. Distribution of MDD of governor-final and governor-initial dependencies.

Figure 8 proves that the MDD of governor-initial dependencies is greater than that of governor-final dependencies in Chinese. We wonder whether it can demonstrate that once a head word appears in Chinese, its complement component will show up later. Figure 9 shows the distribution of the MDDs of governor-final and governor-initial dependencies in 20 languages.[1]

Figure 9 shows that Hungarian is similar to Chinese; both have greater negative MDD. The MDD of governor-final is much higher than that of governor-initial in Japanese and Arabic. However, two MDDs of other languages are close to each other.

[1] These 20 languages are: Chinese (chi), Japanese (jpn), German (ger), Czech (cze), Danish (dan), Swedish (swe), Dutch (dut), Arabic (ara), Turkish (tur), Spanish (spa), Portuguese (por), Bulgarian (bul), Slovenian (slv), Italian (ita), English (eng), Romanian (rum), Basque (eus), Catalan (cat),Greek (ell), Hungarian (hun). Liu (in press) provides more detailed information about these 20 languages (treebanks).

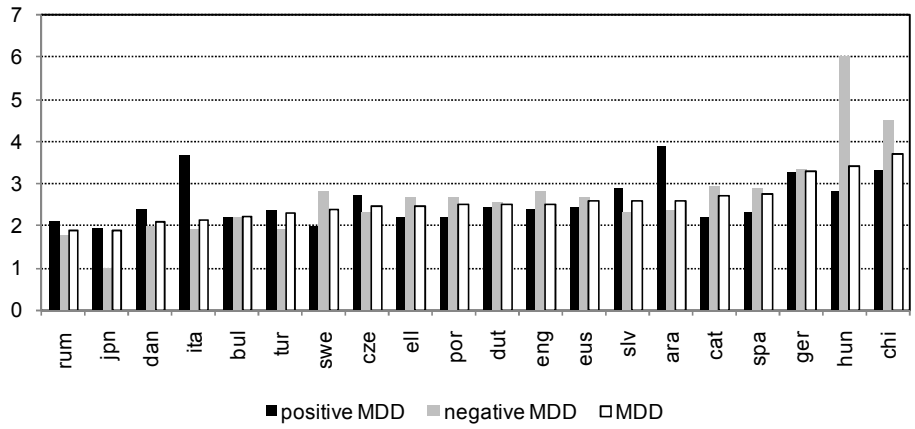

Figure 9. Distribution of MDD of 20 languages.

Therefore, we consider that it is a property relating to individual language. It also raises the fundamental question of whether the mechanisms for processing language vary from one language to another, on which most psycholinguists would have strong views.[2] However, determining what kind of syntactic or linguistic means can cause these differences needs further work.

5. Conclusion

Based on five Chinese dependency treebanks, this paper quantitatively investigates dependency distance and dependency direction in Chinese. The study shows that mean dependency distance of Chinese is 2.84; 40–50% dependencies are between non-adjacent words; Chinese is a mixed language with a governor-final and SV–VO–-AdjN preference, which is consistent with typologists' conclusions obtained through other means. The mean dependency distance of governor-initial dependencies is greater than that of governor-final ones. All of these findings are not only beneficial to investigating Chinese syntactic features based on a dependency grammar formalism, but also useful for the discovery of the universals of human language.

Methodologically, the paper adopts five treebanks with different text genres and annotation schemes as a resource to study syntactic features of a language. This method avoids corpus influences on results so that the conclusions can be more reliable. It is also useful for discovering relations between different syntactic features, as well as universal characteristics of human language through the study of a certain

[2] We thank an anonymous referee for this observation.

language. If relevant treebanks are available, it will be an easy task to apply our method to other languages. In other words, the method has a broad theoretical and cross-linguistic perspective.

The paper finds that adopting treebanks in computational linguistics is beneficial for sharing resources. However, it will be unreliable if we directly use automatically-converted treebanks as a resource of linguistic study. We will further study the influence from treebank annotation schemes on quantitative investigation of syntactical features. Another important problem is how to choose syntactic features to distinguish which ones are individual, and which universal through different treebanks of one language.

ACKNOWLEDGEMENTS

We are grateful to the anonymous reviewers for insightful comments, IR-Lab of HIT for the *HIT* treebank, Joakim Nivre for the *PCT* treebank, the organizers and providers of CoNLL-X 2006 dependency parsing and Academia Sinica for the *Sinica* treebank, and Zhao Yiyi and Guan Runchi for annotating the *Cucn* and *Cucc* treebanks. This work is partly supported by the National Social Science Foundation of China (Grant No. 09BYY024) and the Communication University of China as one of "211" key projects.

REFERENCES

Abeillé, A. (ed.). 2003. *Treebank: Building and using parsed corpora*. Dordrecht: Kluwer.

Best, K.-H. 2006. *Quantitative Linguistik: Eine Annaeherung.* (3rd ed.) Göttingen: Peust & Gutschmidt.

Bod, R., J. Hay and S. Jannedy (eds.). 2003. *Probabilistic linguistics*. Cambridge, MA: MIT Press.

Buch-Kromann, M. 2006. Discontinuous Grammar. A dependency-based model of human parsing and language acquisition. (Unpublished PhD dissertation, Copenhagen Business School.)

Chen, K.-J. et al. 2003. "Sinica treebank: Design criteria, representational issues and implementation". In: Abeillé, A. (ed.). 231–248.

Collins, M. 1996. "A new statistical parser based on bigram lexical dependencies". *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, CA. 184–191.

Cowan, N. 2005. *Working memory capacity*. Hove: Psychology Press.

De Smedt, K., J. Hajič and S. Kübler (eds.). 2007. *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*. December 7–8, 2007. Bergen, Norway.

Gries, S.Th. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge.

Haspelmath, M., M. Dryer, D. Gil and B. Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.

Hudson, R. 1995. Measuring Syntactic Difficulty. <http://www.phon.ucl.ac.uk/home/dick/difficulty.htm> (Date of access: 05 Oct 2009.)
Hudson, R. 2007. *Language networks. The new word grammar.* Oxford: Oxford University Press.
Kakkonen, T. 2005. "Dependency treebanks: Methods, annotation schemes and tools". *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, Joensuu, Finland. 94–104.
Köhler, R. and G. Altmann. 2000. "Probability distributions of syntactic units and properties". *Journal of Quantitative Linguistics* 7(3). 189–200.
Köhler, R., G. Altmann, and R.G. Piotrowski (eds.). 2005. *Quantitative Linguistik. Ein internationales Handbuch* [Quantitative linguistics. An international handbook]. Berlin: Mouton de Gruyter.
Kübler, S., R. McDonald and J. Nivre. 2009. *Dependency parsing*. San Rafael, CA: Morgan and Claypool.
Liu, H. 2007a. "Probability distribution of dependency distance". *Glottometrics* 15. 1–12.
Liu, H. 2007b. "Building and using a Chinese dependency treebank". *Grkg/Humankybernetik*, 48(1). 3–14.
Liu, H. 2008. "Dependency distance as a metric of language comprehension difficulty". *Journal of Cognitive Science* 9(2). 159–191.
Liu, H. 2009a. "Probability distribution of dependencies based on Chinese Dependency Treebank". *Journal of Quantitative Linguistics* 16 (3). 256–273.
Liu, H. 2009b. *Dependency grammar: From theory to practice.* Beijing: Science Press.
Liu, H. In press. "Dependency direction as a means of word-order typology: A method based on dependency treebanks". *Lingua*. doi:10.1016/j.lingua.2009.10.001.
Liu, H., R. Hudson and Zh. Feng 2009. "Using a Chinese treebank to measure dependency distance". *Corpus Linguistics and Linguistic Theory* 5(2). 161–174.
Ma, J. 2007. Research on Chinese dependency parsing based on statistical methods. (Unpublished PhD thesis, Harbin Technology University.)
Marcus, M., B. Santorini and M.A. Marcinkiewicz. 1993. "Building a large annotated corpus of English: The Penn Treebank". *Computational Linguistics* 19(2). 313–330.
Mel'čuk, I.A. 1988. *Dependency syntax: Theory and practice*. Albany: State University Press of New York.
Miller, G. 1956. "The magical number seven plus or minus two: Some limits on our capacity for processing information". *Psychological Review* 63. 81–97.
Ninio, A. 2006. *Language and the learning curve: A new theory of syntactic development.* Oxford: Oxford University Press.
Tesnière, L. 1959. *Eléments de la syntaxe structurale.* Paris: Klincksieck.
Xue, N., F. Xia, F.-D. Chiou and M. Palmer 2005. "The Penn Chinese TreeBank: Phrase structure annotation of a large corpus". *Natural Language Engineering* 11(2). 207–238.

Address correspondence to:
Haitao Liu
Institute of Applied Linguistics
Communication University of China
No. 1 Dingfuzhuang Dongjie
CN-100024, Beijing
P. R. China
lhtcuc@gmail.com