

# Using a Chinese treebank to measure dependency distance

HAITAO LIU, RICHARD HUDSON, and ZHIWEI FENG

## 1 Abstract

2 *This article describes a method for calculating the ‘dependency distance’*  
3 *between the words in a text – i.e. the number of words that separate each*  
4 *word from the word on which it depends syntactically – and reports the*  
5 *results of applying this method to a Chinese treebank. This study shows that*  
6 *Chinese dependencies tend strongly to be governor-final and that the mean*  
7 *dependency distance of words is much higher for Chinese than for other*  
8 *languages that have been studied including English, German and Japanese.*  
9 *It is unclear whether this difference means that Chinese is syntactically more*  
10 *difficult to process.*

11 *Keywords: Dependency syntax; Chinese treebank; dependency distance.*

## 12 1. Introduction

13 A treebank is not only useful for practical projects in computational linguistics  
14 such as training and evaluating parsers, but can also be used as a resource  
15 for quantitatively analyzing the syntactic structures of texts and drawing conclusions  
16 about how humans process the language concerned. In this paper we  
17 report on the dependency patterns in a news Chinese treebank, and discuss  
18 two sets of statistical results involving the direction and length of the dependencies.  
19 The latter measure is ‘dependency distance’, and allows interesting  
20 comparisons between and within languages which we shall discuss. We do  
21 not discuss either the principles of building a treebank (Abeillé 2003), nor  
22 the general relations between treebank and linguistic theories.<sup>1</sup>

23 We shall take it for granted in this paper that the syntactic structure of a  
24 sentence consists of nothing but the dependencies between individual words –  
25 an assumption that is widely accepted not only in computational linguistics  
26 but also in theoretical linguistics and that can be justified independently

27 (Hudson 2007). The following are generally accepted as the core properties  
28 of a syntactic dependency relation (Tesnière 1959; Hudson 1990):

- 29 1. It is a binary relation between two linguistic units.
- 30 2. It is usually asymmetrical, with one of the two units acting as the governor  
31 and the other as dependent.
- 32 3. It is labeled and the type of a dependency relation is usually indicated  
33 using a label on top of the arc linking the two units.

34 A further assumption that we shall make is that the linguistic units related in  
35 this way are single words rather than phrases. As for dependency distance,  
36 this is the linear distance between governor and dependent measured in terms  
37 of the number of words from one to the other. Distance is an important  
38 property of a dependency because of its implications for the cognitive costs  
39 of processing the dependency, so the average dependency distance of a text  
40 is also an important comparative measure for the light that it throws on the  
41 cognitive demands of the language concerned. In this paper, we introduce a  
42 new approach to using a treebank to measure the dependency distance and  
43 dependency directions of a language.

44 Section 2 briefly introduces the dependency syntax of Chinese and the de-  
45 pendency treebank used in this study. The usefulness of dependency distance  
46 is discussed in section 3, and some concepts and our methods for measuring  
47 dependency distance in a treebank are also given in this section. Section 4  
48 presents the results, which we discuss in Section 5.

## 49 **2. Chinese dependency syntax and treebank**

50 Our tentative treebank is built on the news (*xinwen lianbo*) of China Central  
51 Television, a genre which is intended to be spoken but whose style is similar  
52 to the written language. We selected four complete news broadcasts as the  
53 annotated material. The final treebank includes 709 sentences and 17363  
54 word tokens, so the mean sentence length is 24 words.

55 The syntactic analysis was carried out manually. In order to make the  
56 annotation manageable, we recruited two groups of people to annotate the  
57 texts. The first group consisted of 20 undergraduates of Chinese linguistics  
58 who annotated 500 tokens each, and the other group contains a single graduate  
59 of linguistics who did the remaining 10000 tokens. Then two teachers of  
60 Chinese linguistics reviewed the entire corpus.

61 The analysis specifies two kinds of information about each word: its word  
62 class, and its dependency relations to other words defined in terms of a tagset  
63 of dependency types. As far as the word classes are concerned, we used a set  
64 of word classes based on the national standard “Part of speech (POS) tagset

65 for Chinese information processing” (2003) and the widely used “Chinese  
 66 Grammar for middle-school teaching”. Our system includes 13 main classes:  
 67 noun (n), verb (v), adjective (a), adverb (d), pronoun (r), preposition (p),  
 68 numeral (m), classifier (q), conjunction(c), interjection (e), particle (u) and  
 69 onomatopoeia (o) and punctuation (bd). This set excludes some POS tags in  
 70 the national standard set which do not work on the level of syntax, and differs  
 71 from traditional school grammar by giving some functional (particle) words  
 72 an important position in the syntax, because they often play a crucial role by  
 73 determining the dependency relation between two words. Figure 1 shows the  
 74 hierarchy of POS tagset in Chinese.<sup>2</sup>

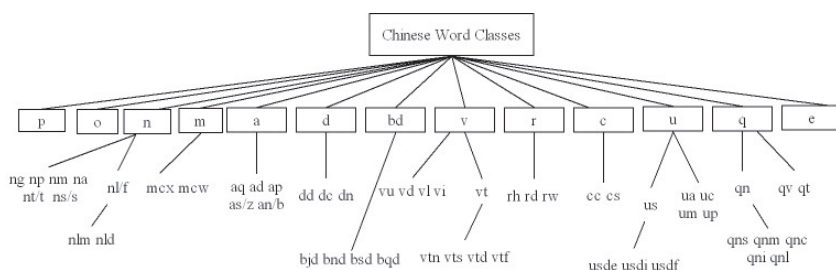


Figure 1. A hierarchy of word classes in Chinese

75 The dependency tagset contains 20 complements and 14 adjuncts. Chinese  
 76 has a few more types of complement than other languages (Maxwell and  
 77 Schubert 1989), because Chinese has to use functional words for signalling  
 78 the grammatical functions which often are morphologically realized in other  
 79 languages. The dependency types are listed in Table 1 and summarised in  
 80 Figure 2.

81 Example (1) is a simple sentence which we can use to illustrate the analysis.

- 82 (1) 这是三个例子。  
 83 ‘This is three classifier example.’ (literal translation)  
 84 ‘These are three examples.’ (regular translation)

85 Table 2 shows the analysis of (1) in terms of this dependency syntax, with  
 86 each word token distinguished by a number which shows the linear order of  
 87 words.

88 This format expresses the properties of each dependency relation, but  
 89 sometimes it is also helpful to construct a connected directed labeled graph  
 90 (Mel’cuk 1988: 23) such as Figure 3.

91 This analysis provides a suitable basis for exploring our two questions about  
 92 the direction and length of dependency relations. However, our treebank is  
 93 still rather small so we also converted the Chinese dependency treebank of

Table 1. Chinese dependency types

Type	Label	Type	Label
Main governor	S	Sentential object	SentObj
Subject	SUBJ	Auxiliary verb	ObjA
Object	OBJ	Coordinating mark	C-
Indirect Object	OBJ2	Adverbial	AVDA
Subobject	SUBOBJ	Verb adjunct	VA
Subject Complement	SOC	Attributer	ATR
Prepositional Object	POBJ	Topic	TOP
Postpositional Complement	FC	Coordinating adjunct	COOR
Complement	COMP	Epithet	EPA
Complement of usde ‘的’	DEC	Numeral adjunct	MA
Complement of usdi ‘的’	DIC	Aspect adjunct	TA
Complement of usdf ‘得’	DFC	Adjunct of sentence end	ESA
Object of Pba ‘把’	BaOBJ	Parenthesis	InA
Plural complement	PLC	Clause adjunct	CR
Ordinal complement	OC	Correlative adjunct	CsR
Complement of classifier	QC	Particle adjunct	AuxR
Construction of Pbei ‘被’	BeiS	Punctuation	Punct

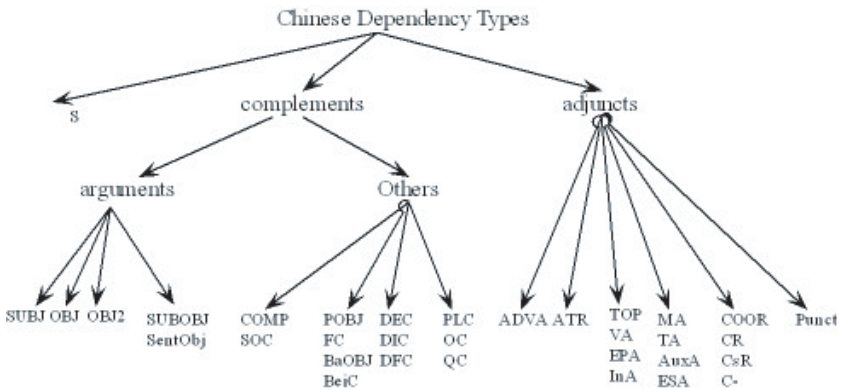


Figure 2. A hierarchy of dependency types in Chinese

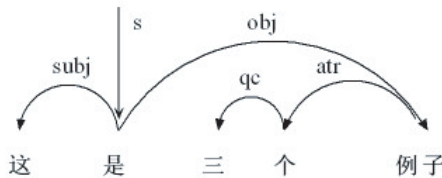


Figure 3. The analysis displayed as a graph

Table 2. *Analysis of a sample sentence*

Dependent			Governor			Dependency
Order	Character	POS	Order	Character	POS	Type
1	这	r	2	是	v	subj
2	是	v	—	—	—	s
3	三	m	4	个	q	qc
4	个	q	5	例子	n	atr
5	例子	n	2	是	v	obj

94 the Harbin Institute of Technology (HIT) into our format. The HIT treebank  
 95 contains 9,996 sentences with 178,467 word tokens, which are mostly taken  
 96 from newspapers – a similar genre selection to our treebank, though the mean  
 97 sentence length is only 17 words, somewhat shorter than the 24 of our corpus.

### 98 3. Measuring dependency distance and dependency direction

99 Dependency distance is the linear distance between governor and dependent.  
 100 The concept was first used in (Heringer et al. 1980: 187), who extracted the  
 101 idea from the depth hypothesis of Yngve on phrase structure grammar (Yn-  
 102 gve 1960, 1996). The term ‘dependency distance’ was introduced in Hudson  
 103 (1995: 16) and defined as “the distance between words and their parents,  
 104 measured in terms of intervening words.”

105 Measuring dependency distance (DD) is useful for:

- 106 1. Predicting syntactic difficulty. If human parsing (i.e. syntactic analysis) of  
 107 a sentence is seen as an incremental procedure for transforming a linear  
 108 string into a tree or graph, a word can be removed from working memory  
 109 only once it can be linked with some other word to form a dependency  
 110 relation. According to Miller (1956), this memory has a limited capacity  
 111 with  $7 \pm 2$  units, which impacts on the human capacity for parsing. This  
 112 question has received wide attention in current cognitive science (Gibson  
 113 1998; Gibson and Pearlmutter 1998; Grodner and Gibson 2005). Gibson  
 114 (1998) hypothesizes “(1) the longer a predicted category must be kept  
 115 in memory before the prediction is satisfied, the greater is the cost for  
 116 maintaining that prediction; and (2) the greater the distance between an  
 117 incoming word and the most local head or dependent to which it attaches,  
 118 the greater the integration cost.” In our case, the greater the dependency  
 119 distance, the harder the processing.
- 120 2. Explaining the mechanisms of children language learning. Ninio (1998)  
 121 shows that children learning both English and Hebrew combine three

words much more easily when dependency distance is the minimum than when one word is separated from the word on which it depends.

3. Designing better parsing algorithms for natural language processing. Collins proves that “the relative order of the words and the distance between them will also strongly influence the likelihood of one word modifying the other.” (1996: 187) Buch-Kromann (2006: 100) also mentions the importance of dependency distance for parsing of a natural language.

Several of these arguments are based on the assumption that DD can be averaged across the words in a text, and that the resulting average DD provides a relevant basis for comparing different texts in a single language, or even for comparing texts in different languages. If two texts in different languages are otherwise comparable – e.g. they are both examples of casual conversation or of scripted news broadcasts – then we may take them as representative of the syntactic patterns in their respective languages and draw conclusions about the languages themselves. Hudson (1995), Hiranuma (1999) and Epler (2004) have used manual analyses of very short texts to calculate the dependency distances in English, Japanese and German. The main innovation of our research project is that the calculation is based on a reasonably large treebank.

How, then, should one calculate the result automatically? Formally, let  $W_1 \dots W_i \dots W_n$  be a word string. For any dependency relation between the words  $W_a$  and  $W_b$ , if  $W_a$  is governor and  $W_b$  is dependent, then the dependency distance (DD) between them can be defined as the difference  $a-b$ ; by this measure, adjacent words have a DD of 1 (rather than 0 as when DD is measured in terms of intervening words). When  $a$  is greater than  $b$ , the DD is a positive number, which means that the governor is after the dependent; when  $a$  is smaller than  $b$ , the DD is a negative number and the governor precedes the dependent. For instance, in the above example “这是三个例子。”, the DD of ‘是—这’ is  $2 - 1 = 1$ ; ‘个—三’ is  $4 - 3 = 1$ ; ‘是—例子’ is  $2 - 5 = -3$ ; ‘例子—个’ is  $5 - 4 = 1$ .

As we shall see below in discussing the direction of dependencies, it is sometimes useful to distinguish the directions of DD with positive and negative numbers, and this is the reason why this project, unlike other projects,<sup>3</sup> always counts the distance between two words as at least 1. However, in measuring dependency distance the relevant measure is the absolute dependency distance (ADD), where the ‘ADD’ between words  $W_a$  and  $W_b$  can be defined as  $ADD = |a-b|$ .

The mean dependency distance (MDD) of an entire sentence can be defined as in (2)

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

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

165 This formula can also be used to calculate the mean dependency distance  
 166 of a larger collection of sentences, such as a treebank; cf. (3).

$$167 \quad (3) \quad MDD(\text{the sample}) = \frac{1}{n-s} \sum_{i=1}^{n-s} |DD_i|$$

168 in this case,  $n$  is the total number of words in the sample,  $s$  is the total  
 169 number of sentences in the sample.  $DD_i$  is the dependency distance of the  
 170  $i$ -th syntactic link of the sample.

171 Another application is to calculate the MDD for a specific type of depen-  
 172 dency relation in a sample, as shown in (4).

$$173 \quad (4) \quad MDD(\text{dependency type}) = \frac{1}{n} \sum_{i=1}^n DD_i$$

174 Here  $n$  as the number of examples of that relation in the sample.  $DD_i$  is the  
 175 dependency distance of the  $i$ -th syntactic link in the set of that dependency  
 176 type.

177 Dependency Distance is not the only important and interesting variable in  
 178 syntactic structure. The other one which we studied is Dependency Direction,  
 179 the contrast between governor-initial and governor-final dependencies. For  
 180 any given dependency, of course, this is a purely qualitative difference, but for  
 181 a collection of dependencies such as those in a sentence or a text it is a quan-  
 182 titative measure according to the relative proportions of the two directions in  
 183 the collection. As explained earlier, our method of measuring Dependency  
 184 Distance allows us to distinguish the two directions according to whether  
 185 the position-number of the governor is higher than that of the dependent  
 186 (governor-final) or lower (governor-initial). Measuring DD of a language can  
 187 more clearly determine if the language is governor-final or governor-initial.  
 188 This idea comes from Tesnière (1959: 22–24, 32–33), who considers such a  
 189 measure useful to studying language typology. Liu (submitted) proposes a  
 190 method based on dependency directions to classify a language typologically,  
 191 and the experiment with 20 languages shows that the dependency direction  
 192 is useful for language typological classification.

#### 193 4. Measuring dependency distance and direction in Chinese

194 There are 17,363 word tokens and 166,54 dependencies in our sample (tree-  
 195 bank), most of which are between adjacent words. In the following report  
 196 our results are normalised to make them comparable to the results of other  
 197 projects (Hudson 1995; Hiranuma 1999; Eppler 2004), where DD is mea-

198 sured in terms of the number of intervening words rather than as the difference  
 199 between the words' position-numbers. For example, adjacent words have a  
 200 DD of 0 in the other projects instead of 1 in our original analysis, and the  
 201 MDD of example (1) is  $(0+0+0+2)/4 = 0.5$ .

202 Table 3 shows the overall frequencies in our tree bank of the various values  
 203 for ADD (the absolute dependency distance, which ignores direction).

Table 3. *The frequency of each distance length in the Chinese treebank*

ADD	Dependencies	Cumulative total (Proportion)
= 0	9377	9377 (56.3%)
= 1	2391	11,768 (70.7%)
= 2	1246	13,014 (78.1%)
= 3	794	13,808 (82.9%)
= 4	583	14,391 (86.1%)
= 5	401	14,792 (88.8%)
= 6	353	15,145 (90.9%)
= 7	282	15,427 (92.6)
= 8	228	15,655 (94%)
= 9	173	15,828 (95%)
≥ 10	826	16,654 (100%)

204 However, ADD interacts strongly with direction. Most dependencies (68.5%)  
 205 are positive (governor-final), but adjacent dependencies are even more likely  
 206 to be governor-final as can be seen in Table 4 and Table 5, where adjacent  
 207 dependencies (where DD = 0) are contrasted with all others. The differences  
 208 in Table 4 are highly significant ( $p < 0.001$ ).

Table 4. *Dependencies in our treebank*

	Positive DD (governor final)	Negative DD (governor initial)	Total
ADD = 0	7563	1815	9378
ADD > 0	3848	3428	7276
	11,411	5243	16,654

Table 5. *Percentage of positive and negative dependencies in our treebank*

	Positive DD (governor final)	Negative DD (governor initial)	
ADD = 0	80.6%	19.4%	56.3%
ADD > 0	52.9%	47.1%	43.7%
	68.5%	31.5%	100%

209 Using the formula mentioned in Section 3, we get the MDD of posi-  
 210 tive/negative and all dependencies in our treebank, which is shown in Table 6.

Table 6. *MDD and its distribution in our treebank*

	Positive DD	Negative DD	All DD
MDD	1.25	3.3	1.89

211 The HIT treebank includes 178,467 word tokens and 168,470 dependencies.  
 212 The related figures are presented in Table 7, Table 8 and Table 9.

Table 7. *Dependencies in HIT treebank*

	Positive DD (governor final)	Negative DD (governor initial)	Total
ADD = 0	68802	16766	85568
ADD > 0	43321	39581	82902
	112123	56347	168470

Table 8. *Percentage of positive and negative dependencies in HIT treebank*

	Positive DD (governor final)	Negative DD (governor initial)	
ADD = 0	80.4%	19.6%	50.8%
ADD > 0	52.3%	47.7%	49.2%
	66.6%	33.4%	100%

Table 9. *MDD and its distribution in HIT treebank*

	Positive DD	Negative DD	All DD
MDD	1.34	2.99	1.89

213 The differences in Table 7 are also highly significant ( $p < 0.001$ ).

214 Our figures also show a strong interaction between dependency distance  
 215 and the individual dependency relations such as *subject* or *object*, which can  
 216 be useful for understanding MDD differences between languages. The data in  
 217 Table 10, calculated using Formula (4), shows the MDD of some dependency  
 218 relations in our Chinese treebank. These figures need more research but they  
 219 show very clearly that some dependencies tend to be much longer than others.

Table 10. *Dependency distances and distribution for distinct dependency types in a Chinese treebank*

Dependency type	Dependencies	MDD <sup>4</sup>
Complements	190	-0.01
Aspect adjunct	217	-0.08
Classifier complement	467	0.11
Attribute	4625	0.55
Subject	1395	1.66
Adverbial	2375	2.44
Object	1673	-2.90
‘Ba’ (把) construction as object	41	4.76
Parenthesis	95	6.8
Sentential object	172	-6.81
Clausal relation	729	-8.96

## 220 5. Discussion

221 Section 4 presents the results of our analysis for two Chinese dependency  
 222 treebanks. In this section, we try to interpret the results and hope to make a  
 223 link between dependency distance and linguistics. During the discussion, we  
 224 will also compare our results with earlier works,<sup>5</sup> although such comparison  
 225 is difficult.

226 Eppler (2004) calculates the dependency distance of English and German.  
 227 After normalization, the distribution of English and German is as shown in  
 228 Table 11 (Eppler 2004: 156–158). Eppler explains the difference in terms of  
 229 the syntactic differences between the two languages. For instance, in German  
 230 the main verb is often at the end of the sentence and word order is freer than  
 231 in English.

Table 11. *Comparison of MDD in samples of English and German speech<sup>6</sup>*

	German	English
MDD	0.87	0.49
ADD = 0	490 (65%)	463 (78%)
ADD > 0	264 (35%)	133 (22%)
Total	754	596

232 Hiranuma (1999: 313) measures the MDD of conversational English and  
 233 Japanese as 0.386 and 0.43, respectively.<sup>7</sup> Japanese is a head-final language,  
 234 which might be expected to have a higher MDD because more dependents  
 235 are separated from the head than in a language like English where the depen-  
 236 dents tend to occur on either side of the head. However, Hiranuma reports

237 that conversational Japanese in fact has a similar MDD to English and sug-  
238 gests, as explanation, that this is because Japanese uses a smaller number  
239 of dependents than English. Hiranuma's experiment also shows that formal  
240 style may have a greater MDD.

241 On the basis of two dependency treebanks,<sup>8</sup> Ferrer i Cancho (2004) cal-  
242 culates that about 50%–67% of the links in sentences are formed between  
243 adjacent words and 16%–25% are formed at distance 1. However, although  
244 his treebanks contain different languages (Romanian and Czech), his method  
245 does not lead to an MDD of the two languages. In contrast, Buch-Kromann  
246 (2006) reports, in the Danish Dependency Treebank,<sup>9</sup> that only 44% of all  
247 dependents are immediately preceded by their governor, though 88% are  
248 fewer than 5 words apart from their governor. Based on the treebanks of 20  
249 languages, Liu (2008a) finds almost 50% of dependency relations are built  
250 between adjacent words.

251 These figures suggest that different languages favour different dependency  
252 distances. On the other hand, dependency distance appears to be positively  
253 correlated with style, with casual speech favouring much shorter distances  
254 than more formal, possibly scripted, speech and writing. This being so, it  
255 is possible that some of these differences between treebanks are due to the  
256 style of the texts concerned rather than to differences between the languages.  
257 However, even if the samples for different languages are taken from different  
258 styles, the numerical differences are striking and deserve more investigation.

259 As for Chinese, the data in Table 3–5 show:<sup>10</sup>

- 260 1. Chinese is basically a governor-final language.
- 261 2. Only 56.3% of dependents are adjacent to their governors, which is a lower  
262 proportion than in English (78%) and German (65%).
- 263 3. The MDD is much greater in governor-initial relations than in governor-  
264 final.

265 According to the studies by Collins (1996) for English, the figures for ADD  
266  $\leq 4$  are 95.6% and for ADD  $\leq 9$  are 99%, which are considerably higher  
267 than for Chinese (Table 3), suggesting a much larger 'tail' of high distances  
268 in Chinese than in English.

269 The interaction of MDD and direction in Chinese also deserves a great  
270 deal more study. According to both Tables 6 and 9, the MDD for governor-  
271 initial dependencies is more than twice as great as that for governor-final  
272 dependencies. This is clearly a feature of Chinese dependencies, but we do  
273 not know either how to generalise to other languages, or to explain it in  
274 Chinese.

275 The above discussion also shows that the MDD of Chinese is much greater  
276 than in English, German and Japanese.<sup>11</sup> If the MDD can be used as a criterion  
277 to estimate the syntactic difficulty, can we say that Chinese is a syntactically  
278 more difficult language than other languages? Or does it show that MDD is

279 only valid in a single language, so we can not use it to compare the syntactic  
280 difficulties of two languages?

281 One way to explore these questions is to control the effect of meaning and  
282 style by looking in detail at a text in one language and its translation into  
283 the other, so we selected a sample text with about 100 words in English,<sup>12</sup>  
284 and compared the dependencies in it and in its published Chinese translation.  
285 The English text contains 106 words and the Chinese translation 108 words.  
286 However in spite of their similar style and meanings, the MDDs of the two  
287 texts were very different: 0.51 for English and 1.42 for Chinese. It is easy  
288 to find structural explanations for these differences; for example, whereas  
289 prepositional phrases follow the modified noun in English, in Chinese they  
290 precede it, which means that the preposition's complement inevitably sepa-  
291 rates it from the modified noun. However, these explanations run counter to  
292 the common functional view that a language's structure adapts to minimise  
293 the processing difficulties of its users. We cannot rule out the possibility that  
294 Chinese makes greater demands of its users than the other languages that we  
295 have studied.

296 The main aim of this paper is not to answer these important theoretical  
297 questions but to show how this kind of research can raise such questions. The  
298 method that we used in the investigations reported here has the following  
299 original features:

- 300 1. It is based on dependency treebanks, which can be built using standard  
301 methods from computational linguistics;
- 302 2. It measures positive and negative dependency distance, i.e. dependency  
303 directions;
- 304 3. It can not only be used to measure the MDD of a sentence, but also to mea-  
305 sure that of a whole language sample or of a specific type of dependency  
306 relation.

307 We leave it to future work to develop these methods to improve our ability to  
308 compare dependency distances across languages and to explore and explain  
309 the reasons for differences of the kind we have revealed here.

### 310 **Bionotes**

311 Haitao Liu is professor of applied and computational linguistics at Commu-  
312 nication University of China (CUC), Beijing. His research interests include  
313 computational linguistics, corpus linguistics and dependency grammar. E-  
314 mail: [lhtcuc@gmail.com](mailto:lhtcuc@gmail.com)

315 Richard Hudson is an emeritus professor of linguistics in the Department  
316 of Phonetics and Linguistics at University College London.

317 Zhiwei Feng is professor at Communication University of China (CUC),  
318 Beijing.

## 319 Notes

- 320 \* We thank two anonymous reviewers for their insightful comments, Zhao Yiyi for annotat-  
321 ing the treebank. This work is partly supported by Communication University of China  
322 as one of 211 key projects.
- 323 1. Six consecutive TLT (Treebanks and Linguistic Theories) workshops reveal that there  
324 is close relationship between treebanks and linguistic theories. The first “Treebanks and  
325 Linguistic Theories” workshop was held in Sozopol, Bulgaria in 2002; the second in Vaxjo,  
326 Sweden in 2003; the third in Tuebingen, Germany in 2004; the fourth in Barcelona, Spain  
327 in 2005; the fifth in Prague, Czech in 2006 and the sixth in Bergen, Norway in 2007.
  - 328 2. Liu and Huang (2006) includes a detailed explanation of the subclass tags.
  - 329 3. Collins (1996) and Ferrer i Cancho (2004) also define dependency distance as the order  
330 difference between governor and dependent, but they and other colleagues (Hudson 1995;  
331 Hiranuma 1999; Eppler 2004) do not distinguish the direction of a DD and only use the  
332 absolute value of the difference.
  - 333 4. Here the signs (–) or (+) to indicate if the dependent is before or after the governor. For  
334 instance, if MDD has a value –2.5, which means that the dependency type is governor-  
335 initial and there are approximately 1.5 words between the governor and the dependent.
  - 336 5. We limit our discussion to dependency grammar formalism, although there are also some  
337 works from cognitive science, which often are under phrase structure formalism.
  - 338 6. While Eppler and Hiranuma calculate a MDD of a sentence with  $n$  words, they use the  $n$   
339 as the divisor. The difference is very small. In our case, the MDD using  $n$  is 1.81, using  
340  $n - 1$  is 1.89.
  - 341 7. English: 1035 words; Japanese: 2117 words.
  - 342 8. Romanian with 21,275 words and 2340 sentences, Czech with 563,067 words and 31,701  
343 sentences.
  - 344 9. 100,200 words, 5540 sentences.
  - 345 10. Liu (2008b) has similar conclusions based on 5 Chinese treebanks with different genres  
346 and annotation schemes.
  - 347 11. According to Liu (2008a), Chinese has also the greatest dependency distance in 20 lan-  
348 guages, which are Chinese, Japanese, German, Czech, Danish, Swedish, Dutch, Arabic,  
349 Turkish, Spanish, Portuguese, Bulgarian, Slovenian, Italian, English, Romanian, Basque,  
350 Catalan, Greek, Hungarian.
  - 351 12. The text is extracted from Pinker’s “The Language Instinct” (Perennial Classic, 2000). De-  
352 pendency analysis at <http://www.phon.ucl.ac.uk/home/dick/enc/pinker.htm>. The Chinese  
353 translation is from the Chinese version of this book (Shantou University Press, 2004).

## 354 References

- 355 Abeillé, Anne (ed.)  
356 2003 *Treebank: Building and using parsed corpora*. Dordrecht: Kluwer.
- 357 Buch-Kromann, Matthias  
358 2006 *Discontinuous grammar. A dependency-based model of human parsing and lan-  
359 guage acquisition*. Copenhagen: Copenhagen Business School Dr. ling.merc.  
360 dissertation.
- 361 Collins, Michael  
362 1996 A new statistical parser based on bigram lexical dependencies. *Proceedings of  
363 the Association for Computational Linguistics* 34. 184–191.

- 364 Eppler, Eva  
365 2004 *The syntax of German-English code-switching*. London: University College  
366 London unpublished PhD.
- 367 Ferrer i Cancho, Ramon  
368 2004 Euclidean distance between syntactically linked words. *Physical Review E* 70.  
369 056135.
- 370 Gibson, Edward  
371 1998 Linguistic complexity: locality of syntactic dependencies. *Cognition* 68(1).  
372 1–76.
- 373 Gibson, Edward & Neal J. Pearlmutter  
374 1998 Constraints on sentence comprehension. *Trends in Cognitive Sciences* 2(7).  
375 262–268.
- 376 Grodner, Daniel & Edward Gibson  
377 2005 Consequences of the serial nature of linguistic input for sentential complexity.  
378 *Cognitive Science* 29(2). 261–290.
- 379 Heringer, Hans-Jürgen, Bruno Strecker & Rainer Wimmer  
380 1980 *Syntax: Fragen-Lösungen-Alternativen*. Munich: Wilhelm Fink Verlag.
- 381 Hiranuma, So  
382 1999 Syntactic Difficulty in English and Japanese: A textual study. *UCL Working  
383 Papers in Linguistics* 11. 309–322.
- 384 Hudson, Richard  
385 1990 *English Word Grammar*. Oxford: Blackwell.
- 386 Hudson, Richard  
387 1995 Measuring syntactic difficulty. Unpublished paper.  
388 <http://www.phon.ucl.ac.uk/home/dick/difficulty.htm>
- 389 Hudson, Richard  
390 2007 *Language networks: The new Word Grammar*. Oxford: Oxford University Press.
- 391 Liu, Haitao  
392 2008a Dependency distance as a metric of language comprehension difficulty. *Journal  
393 of Cognitive Science* 9(2). 159–191.
- 394 Liu, Haitao  
395 2008b A quantitative study of Chinese structures based on dependency treebanks.  
396 *Yangtze River Academic* 3. 120–128.
- 397 Liu, Haitao  
398 submitted. Dependency direction as an index of linguistic typology: A method based on  
399 dependency treebanks.
- 400 Liu, Haitao & Wei Huang  
401 2006 A Chinese dependency syntax for treebanking. In *Proceedings of the 20th  
402 Pacific Asia Conference on Language, Information and Computation*, 126–  
403 133. Wuhan, Beijing: Tsinghua University Press.
- 404 Maxwell, Dan & Klaus Schubert  
405 1989 *Metataxis in practice: Dependency syntax for multilingual machine translation*.  
406 Dordrecht: Foris.
- 407 Mel'cuk, Igor A.  
408 1988 *Dependency syntax: Theory and practice*. Albany: State University Press of  
409 New York.
- 410 Miller, George  
411 1956. The magical number seven plus or minus two: Some limits on our capacity for  
412 processing information. *Psychological Review* 63. 81–97.

- 413 Ninio, Anat.  
414 1998 Acquiring a dependency grammar: The first three stages in the acquisition of  
415 multiword combinations in Hebrew-speaking children. In G. Makiello-Jarza,  
416 J. Kaiser & M. Smolczynska (eds.). *Language acquisition and developmental*  
417 *psychology*. Crakow: Universitas.
- 418 Tesnière, Lucien  
419 1959 *Eléments de la syntaxe structurale*. Paris: Klincksieck.
- 420 Yngve, Victor.  
421 1960 A model and a hypothesis for language structure. *Proceedings of the American*  
422 *Philosophical Society* 104(5). 444–466.
- 423 Yngve, Victor  
424 1996 *From grammar to science: New foundations for general linguistics*. Amsterdam  
425 & Philadelphia: John Benjamins.