The effects of sentence length on dependency distance, dependency direction and the implications—Based on a parallel English–Chinese dependency treebank

Jingyang Jiang a, Haitao Liu a,b,*

a Department of Linguistics, Zhejiang University, Hangzhou 310058, China
b Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China

A R T I C L E   I N F O
Article history:
Received 5 September 2014
Received in revised form 31 March 2015
Accepted 6 April 2015
Available online

Keywords:
Dependency distance
Sentence length
Dependency direction
Working memory
Probability distribution

A B S T R A C T

Dependency distance is closely related to human working memory capacity, but is also influenced by other non-cognitive factors. Studies of dependency distance contribute to the understanding of the universalities and peculiarities of languages as well as human cognitive processes in language. Forty two sentence sets were selected from a parallel English–Chinese dependency treebank to examine the progressive properties of dependency distance with the change of sentence length in the two languages. It was found that: (1) the probability distribution models of dependency distance of both languages are not affected by either sentence length or the type of language; (2) the quantity of adjacent dependencies in the two languages are identical, but the quantity of adjacent dependencies of Chinese fluctuates within a limited range, while that of English shows a falling tendency; (3) the mean dependency distances (MDDs) of Chinese are always higher than those of English, and both MDDs show slight ascending trends; (4) compared with dependency distance, dependency direction is a more reliable metric for language classification. These findings suggest that: (1) the universal cognition mechanism may be the major factor affecting the general traits of dependency distance, while language-related factors such as sentence length may affect certain traits of dependency distance; and (2) Chinese taxes working memory more than English.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

Dependency distance (DD, or dependency length) refers to the linear distance between two linguistic units having a syntactic relationship within a sentence (Heringer et al. 1980; Hudson, 1995). If the understanding and analysis of a sentence is comparable to the process of converting a linear string of words into a dependency tree (graph), then a word can be removed from our working memory only when this word encounters its head and forms a dependency relationship (or a more complex concept) (Ferrer-i-Cancho, 2004; Hudson, 2010; Liu, 2008). The linear distance between two words with syntactic relationship is thus restrained by human working memory capacity. A greater dependency distance may overload a human’s working memory, and make the sentence difficult to understand. Dependency distance of a sentence reflects the
difficulty degree of analyzing a given sentence at a syntactic level. The greater the dependency distance, the more difficult it is to analyze the sentence structure (Gibson, 1998; Gibson and Pearlmutter, 1998; Hiranuma, 1999; Liu, 2008).

The analysis of dependency distance contributes to the understanding of the universalities and peculiarities of human cognitive processes in language as well as language itself. If dependency distance is related to human working memory capacity, then its distribution in human language should abide by certain general laws. Human working memory capacity is believed to be similar and limited (Miller, 1956; Cowan, 2001), which defines universality. However, dependency distance also reflects the features of word order with syntactic relationship within a sentence, and word order is an important metric in modern language typology (Song, 2012), thus dependency distance may also exhibit some language-specific uniqueness.

The Depth Hypothesis (Yngve, 1960) is a hypothesis on the relationship between working memory capacity and the complexity of language structure comprehension. Later it was introduced by Heringer et al. (1980) into dependency grammar, which enables one to study the relationship between the two (the complexity of language structure comprehension and working memory capacity) under the framework of dependency syntax. Humans are believed to have adopted an incremental parsing strategy when they comprehend sentences. Words that fail to form structures will be kept in the working memory temporarily. But, because working memory capacity is limited, if the stored words overload the working memory, a breakdown of comprehension may occur (Covington, 2003).

From a psycholinguistic perspective, the syntax analysis model based on working memory is well-grounded (Jay, 2004; Levy et al. 2013), but in the field of cognitive science, this issue is usually determined by the relationship between the difficulty level of understanding the sentence structure and the linear order (DD under dependency grammar) of the words with syntactic relations (Gibson, 1998; Gibson and Pearlmutter, 1998; Gibson, 2000; Grodner and Gibson, 2005; Temperley, 2007; Liu, 2008; Gildea and Temperley, 2010; Fedorenko et al. 2013). Several memory-based or distance-based theories have been proposed, including Early Immediate Constituents (EIC) (Hawkins, 1994), Minimize Domain (MiD) (Hawkins, 2004), Dependency Locality Theory (DLT) (Gibson, 2000), etc. These theories embody the hypothesis that longer dependencies are more difficult to process. Some of these operational theories have become a component of the syntactic synergetic model proposed by Köhler (2012) and are conducive to our understanding of human languages from the perspective of system theory.

Mean dependency distance (MDD) of a sentence is a good predictor of syntactic difficulty as found by the analysis of the dependency distance of sentences which present syntactic difficulty in psycholinguistic experiments (Liu, 2008; Hudson, 1996). Similar conclusions are based on sentences with special structures in languages such as English, German, and Dutch (Lin, 1996) and Japanese (Hiranuma, 1999). There is a general tendency to minimize the mean dependency distance in human languages (Ferrer-i-Cancho, 2004, 2006; Liu, 2007, 2008), but not in random languages (Ferrer-i-Cancho, 2004; Liu, 2007, 2008; Gildea and Temperley, 2010). Mean dependency distance is also proved below chance in various languages (Ferrer-i-Cancho and Liu, 2014).

The fact that dependency distance shares some common characteristics is also evidenced by the findings that the probability distribution of dependency distance is found to abide by certain models (Liu, 2007; Ferrer-i-Cancho and Liu, 2014). Even though DDs of different languages are subject to similar human cognitive mechanisms, they have specific differences (Ferrer-i-Cancho, 2004; Liu, 2008; Gildea and Temperley, 2010). Do these differences suggest that languages may not be equally demanding concerning working memory, i.e. some languages may tax working memory more than others because languages may differ at the level of dependency distance? For instance, it has been found that the MDD of Chinese is at least twice as great as that of English (Liu, 2008; Liu et al. 2009a). Do other corpora in Chinese and English show the same differences? If they do, is it because people’s working memories are different in the two languages (Hudson, 2009) or is it because the two languages have their particular traits in terms of their sentence length and syntactic dependency, or both? To answer these questions, we will examine the related properties of dependency distance of Chinese and English, and discuss their implications.

Dependency distance can be affected by sentence length, the type of text, and the annotation scheme. While calculating dependency distance, some studies mix the sentences with varying length (Hudson, 1995; Hiranuma, 1999; Temperley, 2007; Liu, 2007, 2008; Gildea and Temperley, 2010); some others control the sentence length but do not control the genre or style of the texts (Ferrer-i-Cancho, 2004); still others control neither of the sentence length nor the genre (Liu, 2008; Gildea and Temperley, 2010). All this could result in data deviation, yield distorted results and thus lead to unreliable findings related to dependency distance. First, a different global MDD from sentences of varying length may not be fine-tuned enough to reflect the peculiarities of a language. Second, from a network theoretic perspective, the MDD of a sentence is believed to be related to sentence length (Ferrer-i-Cancho, 2013). Third, the differences in MDD (even with the control of SL) may simply be due to the different genre or syntactic annotation scheme of the languages in question. Therefore, it is more desirable to use a parallel corpus with controlled sentence length, same genre and similar syntactic annotation schemes as we did in the present study in order to compare the properties of dependency distance of English and Chinese at great length. It is hoped that this study may provide some tentative answers to the afore-mentioned questions by focusing on the following four more concrete and technical questions:

---

1 Ferrer-i-Cancho (2004) shows that sentences have a mean dependency length that is below chance (below that of a random language) but greater than the minimum possible because of the restriction of grammar.
1) What is the distribution of dependency distance of natural texts of both English and Chinese with different sentence lengths? Do they follow certain laws or models?

2) Will the quantity of adjacent dependencies change with the increase of sentence length in both languages?

3) Will the mean dependency distance become longer with the increase of sentence length in both languages?

4) Will the percentage of dependency direction change with the increase of sentence length in both languages?

Question 1 is intended to find the universalities of the two languages, and the other three the possible peculiarities. All the findings from these four questions will help answer the ultimate question whether English or Chinese taxes one’s working memory more at the level of dependency distance. Details of the research methods and language materials will be introduced in Section 2. Section 3 will present the results and discuss the probability distribution of dependency distance of both English and Chinese, the relationship between the quantity of adjacent dependencies and sentence length, the relationship between mean dependency distance and sentence length, and the relationship between dependency direction and sentence length. Conclusions and implications are discussed in the last section.

2. Methods and materials

The concept of (dependency) distance is often used in the syntactic analysis framework with phrases or dependency relations as its basic constituents. The present paper uses the syntactic analysis framework of dependency grammar in which sentence structure is analyzed using the dependency relations between words in a sentence (Tesnière, 1959; Hudson, 2007, 2010; Nivre, 2006; Liu, 2009). A dependency relation has three core features: binary, asymmetry, and labeledness.

Based on these three features, one can construct a syntactic dependency tree or directed dependency graph as a representation of a sentence. We chose directed acyclic graphs to illustrate dependency structure in Fig. 1, which presents a dependency analysis of the sentence I actually live in New York.

In Fig. 1, all the words in the sentence are connected by grammatical (or dependency) relations. In each pair of connected words, one is called the dependent and the other is called the head. The labeled arc extends from the head to the dependent. The number below each word indicates its position in the sentence, which is used in computing dependency distance. Many dependency analyses do not use such numbers.

The term “dependency distance” introduced by Hudson (1995) is defined as “the distance between words and their parents (or heads), measured in terms of intervening words.” For computing dependency distances for large corpora, Liu et al. (2009a) proposed a method for measuring the MDD of a sentence with a sample of a treebank (a corpus with syntactic annotation). Formally, let W₁...Wₙ be a word string. For any dependency relation between the words Wₓ and Wᵧ (1 ≤ x, y ≤ n), if Wₓ is a head and Wᵧ is its dependent, then the dependency distance between them is defined as the difference x−y; by this measure, the DD of adjacent words is 1. When x is greater than y, the DD is a positive number, which means the head follows the dependent; when x is smaller than y, the DD is a negative number and the head precedes the dependent. However, in measuring DD the relevant measure is the absolute value of DD.

The MDD of an entire sentence can be defined as:

\[
MDD(\text{the sentence}) = \frac{1}{n-1} \sum_{i=1}^{n-1} |DDᵢ|
\]  

(1)

Here n is the number of words in the sentence and DDᵢ is the dependency distance of the i-th syntactic link of the sentence. Usually in a sentence there is one word (the root verb) without a head, whose DD is defined as zero.

This formula can also be used to calculate the MDD of a larger collection of sentences, such as a treebank:

\[
MDD(\text{the sample}) = \frac{1}{n-s} \sum_{i=1}^{n-s} |DDᵢ|
\]  

(2)

In this case, n is the total number of words in the sample, s is the total number of sentences in the sample and DDᵢ is the dependency distance of the i-th syntactic link of the sample.
Thus, in the sentence “I actually live in New York” (Fig. 1), the DD of “I” is 2, which is obtained by subtracting 1 from 3. So a series of dependency distances can be obtained as follows: 2 1 0 1 1 and this sentence has five dependencies: three dependencies with $DD = 1$, one with $DD = 2$, and one with $DD = 0$. Using Formula (1), the MDD of this sentence is $5/4 = 1.25$.

To avoid the possible disturbing effects of sentences with mixed SLs, of different genres, and of different annotating schemes, a small-sized parallel English–Chinese dependency treebank (Li, 2012) was chosen. There are two reasons for using Chinese and English as the investigated languages, one for convenience, and the other is that Chinese and English belong to two different language families: English, the Indo-European family, and Chinese the Sino-Tibetan family. So the commonalities (cognition-restrained) found in these two languages may be more convincing. Moreover, more peculiarities (language-restrained) are likely to be found. The texts are selected from news items in China Daily, which is written in English either by English native speakers or Chinese journalists but edited by English native speakers. The corresponding Chinese texts are translated from the English version by Chinese native speakers. Therefore, the texts, whether in English or Chinese, are authentic and idiomatic. The treebank we used has 763 English sentences, 20,067 words, with an average sentence length of 26.3, and 882 Chinese sentences, 21,172 words, with an average sentence length of 24.

A thorough comparison and alignment was done on the annotating schemes of English and Chinese before annotating the parallel corpus to eliminate the influence of different annotation schemes.

The graphic sentence length distribution of our English–Chinese parallel treebank and the number of sentences for each length are shown in Fig. 2.

It can be seen from Fig. 2 that in our parallel treebank, 20-word-long sentences are in a majority. To warrant sufficient sampling for all the sentence lengths included and an homogenous number of sentences per sentence length, 10 sentences (G. Altmann, personal communication) were randomly selected for each sentence length (from 10 to 30 words). Sentences whose length is shorter than 10 or longer than 30 are not included in our sample. These 420 sentences (210 in English and 210 in Chinese) make up 42 sets (a set has 10 sentences with the same length, either in English or Chinese), whose SL ranges progressively from 10 to 30 words. This affords a more accurate and dynamic study of (M)DDs when the SLs are on a gradual rise, giving us a clearer interpretation of the dynamic relationship between SL and DD. For instance, if one set has shorter sentences, and another has longer sentences, what effects does this difference have on other quantitative measures of dependency distance (syntactic difficulty)? Will the two sets whose SL is 10 and 30 respectively have the same MDD? What about the quantitative measures of dependency distance of the same length but different languages? These questions may be answered in Section 3, which presents the results derived from these 42 sets, followed by detailed discussions.

3. Results and discussions

3.1. The probability distribution of dependency distance

If dependency distance can indicate how human working memory constrains language comprehension and production, and if most people have similar working memory capacity, it is probable that the distribution of dependency distance of human languages will follow certain regularity. Probability distribution of dependency distance of natural texts can contribute to our understanding of its general features.

The quantities of various dependency relations related to DD of each set were computed with quantitative linguistic software of Altmann-Fitter to determine the probability distribution models suitable for each set’s DD.²

---

From only the values of the coefficient of determination $R^2$ among the 42 sets in English and Chinese, the DD of these sets reveal the following probability distribution:

- Right truncated Kemp2 ($a$),
- Right truncated Salvia-Bolinger ($a$),
- Right truncated Waring ($b, n$),
- Right truncated zeta ($a; R = x\text{-max}$),
- Right truncated Waring ($b, n$),
- Right truncated negative binomial ($k, p; R = x\text{-max}$),
- Right truncated modified Zipf-Alekseev ($a, b; n = x\text{-max}, a$ fixed).

The mean values of $R^2$ of the 42 sentence sets with suitable distribution models were calculated and tabulated as shown in Table 1.

Table 1 data indicate that the probability distribution of the dependency distance shows some regularities apparently not influenced by sentence length or the type of language. Even though among all the figures of distribution, the index of Right truncated zeta is not conclusive, it is acceptable. These results are similar to Liu’s (2007), but he pooled all the sentences of different lengths together. Because Ferrer-i-Cancho’s (2004) study reveals that the DD of sentences with fixed length fits exponential distribution, we fit exponential distribution to our 42 sets. The results (Chinese $R^2 = 0.979$, English $R^2 = 0.989$) are consistent with Ferrer-i-Cancho’s finding that exponential distribution can well fit the distribution model of dependency distance.

Thus, it can be concluded that the distribution model of DD is independent of sentence length ($10 \leq SL \leq 30$), and the distribution models of DD of all the 21 sets with progressive SL (from 10 to 30) in both English and Chinese are identical. Within our present range of SL, it is difficult to distinguish a distribution model of the DD of sentences with a fixed SL and that of the DD of sentences with a mixed SL. Admittedly, if we further expand our range of SL, the distribution of DD may be more like exponential distribution as pointed out by Ferrer-i-Cancho (2004), and needs to be verified with a larger treebank with more diverse languages. Our present research focuses on the similarities of the DD distribution models of different SLs and of different languages, but not the one distribution model that fits our sample best. We will examine why the DD distributions of different SLs and languages have such similarities from the forms of these distribution models.

Fig. 3 presents the fitting of Right truncated zeta to the DD of twenty-word-long sentences, 10 sentences in English and 10 in Chinese.

Table 1 indicates that the chosen language sample simultaneously fits several probability distribution models. Do these models show any significant differences in form? Fig. 4 presents fitting graphs of several probability distributions (cf. Table 1) of 20-word-long English sentences.

---

Table 1

<table>
<thead>
<tr>
<th></th>
<th>Modified Zipf-Alekseev</th>
<th>Salvia-Bolinger</th>
<th>Negative Binomial</th>
<th>Kemp2</th>
<th>Waring</th>
<th>Zeta</th>
</tr>
</thead>
<tbody>
<tr>
<td>English $R^2$</td>
<td>0.949</td>
<td>0.991</td>
<td>0.993</td>
<td>0.991</td>
<td>0.994</td>
<td>0.994</td>
</tr>
<tr>
<td>Chinese $R^2$</td>
<td>0.995</td>
<td>0.992</td>
<td>0.994</td>
<td>0.992</td>
<td>0.992</td>
<td>0.988</td>
</tr>
</tbody>
</table>

---

Fig. 3. Fitting of a right truncated zeta to the DD of twenty-word-long sentences (English on the left and Chinese on the right).

---

3 For more technical information about these distributions, refer to Wimmer and Altmann (1999) and Altmann-Fitter (2013). For more theoretical discussions about DD and its minimization, refer to Ferrer-i-Cancho (2004, 2013), and Ferrer-i-Cancho and Liu (2014).
Figs. 3 and 4 indicate that distribution models exhibit great similarities in form whether it is the same distribution of the same sentence length and different languages, or the different distributions of the same language.

Data from Fig. 4 indicate that all the distributions demonstrate great long tail similarities. Among them, regardless of which distribution, the number of the dependency relation between adjacent words (the DD is 1) is highest. Dependency relations whose DD is 1 have a critical role in forming such asymmetrical distribution. The number of the dependency relations between adjacent words is one of the important factors influencing the MDD of a sentence or a text (Liu, 2008). What then, is the relationship between the number of adjacent dependencies and sentence length? Will the quantity of adjacent dependency relations increase with the increase of sentence length? The answer to this question helps explain why the English and Chinese data share similar probability distribution models.

### 3.2. The relationship between the quantity of adjacent dependencies and sentence length

About half of the dependency relationships are formed between neighboring words in human languages based on research with twenty languages (Liu, 2008). The percentage is 74.2 for English language (Collins, 1996). In a study based on a Danish dependency treebank, Buch-Kromann (2006) reported 44% of all dependents are immediately preceded by their head, and 88% are separated from their head by fewer than 5 words. Ferrer-i-Cancho (2004) found that for Romanian and Czech treebanks about 50%–67% of the links in sentences are formed between adjacent words and 16%–25% are formed at a distance of 2. Liu et al. (2009b) found that the percentages of adjacent dependencies of Chinese are between 47.9 and 56.6, the average being 53.06, using five Chinese treebanks with different annotating schemes, genres and average SLs. Because forming dependency relations among adjacent words produces the lowest cognitive load, it is one of the major reasons of DD minimization of human languages. This has been verified not only by previous studies (Ferrer-i-Cancho, 2004; Liu, 2008), but also supported by related theoretical analysis (Ferrer-i-Cancho, 2004, 2013).

We found the average percentage of adjacent dependencies of English is 61.7, and that of Chinese is 59.6, which indicates no significant difference. Both changes are kept within a stable range from 55% to 67% for English, and from 55% to 64% for Chinese. This striking similarity helps explain why the Chinese and English data of DD both share similar probability distribution models. It also suggests that the working memory capacity may play a key role in restraining the DD. Fig. 5 indicates that even though the average percentages of adjacent dependencies of the two languages do not show obvious differences, their relationships with SL differ greatly. The use of sentences whose length increases one by one enables us to observe the progressive change of the percentage of adjacent dependencies with the increase of SL. With the increase of sentence length, in Chinese, the percentages of adjacent dependencies fluctuate only within a certain range, instead of showing synchronized change (Pearson correlation = 0.022, P-Value = 0.925), while in English, the percentages of adjacent dependencies show a falling tendency (Pearson correlation = -0.712, P-Value = 0). This indicates the number of adjacent dependencies not only changes with SL, but also with the type of language. This may be defined as peculiarity of the language, a conclusion consistent with the observations by Liu (2008) and Liu and Xu (2012), where they used sentences with mixed lengths. Again, this suggests that mixing the sentence lengths will not have significant influence over the overall ratio of adjacent dependencies for our range of SL. But to the ratio of adjacent dependencies from sentences of different lengths and different languages, the mixing of SL might raise some problems, because it may obscure the dynamic attributes of adjacent dependencies with the change of SL in different languages.

Because the number of adjacent dependencies is related to sentence length, then the mean dependency distance of one sentence or one set of sentences is bound to be related to sentence length. What, exactly, is this relationship?
3.3. The relationship between sentence length and mean dependency distance

Longer sentences are presumed to have longer dependency distance, and also longer mean dependency distance, since it is the sentence length that determines MDD. A long sentence is the precondition of a long dependency distance. One could have a long sentence where all the individual dependencies are short, but a high MDD requires a high SL. The dependency distance of a random language, which does not follow syntactic rules, will be more readily influenced by sentence length, but the grammar and cognitive mechanisms will cause the dependency distance of natural languages to become as short as possible (minimization) (Liu, 2007, 2008; Gildea and Temperley, 2010). The relationship between sentence length and dependency distance is assumed to be linear, i.e. $y = a + bx$, where $x$ stands for SL, $y$ stands for the corresponding MDD. Does this theoretical assumption best describe the actual language use?

The linear relationship between MDD and SL of English and Chinese is presented in Fig. 6.
MDDs of every set of Chinese sentences are higher than those of English, and both MDDs gradually rise with the increase of SL chosen for our present study. Among them, the linear fitting formula for English is $y = 1.6521 + 0.0365x$ ($R^2 = 0.793$, Pearson correlation = 0.891, P-Value = 0), and Chinese is $y = 2.0078 + 0.0361x$ ($R^2 = 0.5625$, Pearson correlation = 0.752, P-Value = 0). This result is similar to Ferrer-i-Cancho’s (2004) which is based on a Romanian treebank, but not fully comparable, because he did not exclude sentences whose lengths are either shorter than 10 or longer than 30. The linear formula he obtained is $<d> = 1.163 + 0.039n$, where $<d>$ stands for MDD, and n for SL. From the linear functions of these three languages, the slope $b$ is nearly identical, and parallel, though there is an obvious difference in “$a$”. This difference in “$a$” is the major factor that has caused the differences in the MDDs among different languages. The similarity of “$b$” demonstrates that even though MDDs of both languages studied increased with SL, the acceleration speed is equally slow, and the small value of “$b$” indicates that the MDDs of both languages change only within a small range.

Even though the linear function $y = a + bx$ can reflect the relationship between MDD and SL, considering the fact that the MDD can only increase slowly with SL, it may be more informative to adopt a nonlinear function if one intends to study all the sentences with various lengths, not limited only to our present sentence lengths. Ferrer-i-Cancho and Liu (2014) reported changes in the relationship between DDs of different sentence lengths and SL from four corpora: Basque, Catalan, Spanish and Czech. Compared with that data, it is clear that the range of SL chosen for our present study produced the relatively more concentrated MDD changes, and is more useful for the discovery of new laws of language. In the four corpora, with the decrease of sentence quantity, DDs of some sentences have violent fluctuations, which are unfavorable for discovering language laws and should be avoided for the present purpose. Violent changes cause the fitting results of linear function to the whole treebank to be less desirable. Ferrer-i-Cancho and Arias (2013) fitted several nonlinear functions to a Catalan treebank, and their studies suggest an almost power-law dependency between MDD and SL. We fitted power function to our sample, and the results are presented in Fig. 7. Compared with Fig. 6, the difference in “$a$” between English and Chinese is still obvious, but the similarity in “$b$” is slightly lower than that of linear function.

The coefficient of determination $R^2$ in Chinese (Fig. 7) is slightly higher than that in Fig. 6, while that in English is slightly lower.

In addition to the linear function results in Fig. 6 and a power law function results in Fig. 7, fitting of an exponential function between SL and MDD (Fig. 8) was also done on our sample.

The coefficient of determination $R^2$ of Chinese in Fig. 8 is slightly higher than that in Fig. 6 but slightly lower than that in Fig. 7. Even though the $R^2$ of English in Fig. 8 falls between the $R^2$’s in Figs. 6 and 7, the result is just the opposite of Chinese: it is higher than that in Fig. 7 and lower than that in Fig. 6. With our present sample, it is hard to determine which of these three functions is best suited to reflect the relationship between SL and MDD, but Figs. 6–8 all confirm that the increase of MDD with SL is very slow in both English and Chinese.

What factors are responsible for the similarities and slow change in MDD change in English and Chinese? In Section 3.2, it was demonstrated that no matter how long the sentences are, more than half of the dependency relations are always formed between two adjacent words. Because the DD of adjacent words is 1, this reduces the MDD, and also reduces a human’s working memory load. According to Ferrer-i-Cancho (2004), in a random ordering sentence, its MDD = (SL + 1)/3. Based on this, we calculated the baseline MDDs of the sentences with 21 different lengths, and they fall between 3.67 and 10.33, with
an average MDD of 7. Evidently, the baseline MDD of random ordering sentences is much higher than MDDs of the two languages studied, which reflects that the actual MDD of human languages is restrained/optimized by working memory or grammar. This constraint of working memory capacity and the DD minimization effect both cause the MDD to fluctuate within a very narrow band. These joint factors contribute to the slow increase of the mean dependency distance with the increase of sentence length.

The MDDs of English and Chinese in our sample (21 sets for each language) are different, English being 2.053, Chinese being 2.406. From the trajectory (Figs. 6–8) of the change in the MDDs, the highest of MDD in English is about equal to the lowest of MDD in Chinese, suggesting that the MDDs of 21 Chinese sentence sets, as a function of sentence length, are all above those of English. Furthermore, the MDD of Chinese keeps increasing even though it was already higher than that of English. Since parallel treebank is used, these differences in MDD cannot be attributed to the differences in genre. These figures of MDDs suggest that 1) Chinese taxes one’s working memory more than English; 2) English is more optimized than Chinese at the level of dependency distance; and 3) the same figure of DD may mean differently for Chinese and English because the average word length of Chinese is shorter than that of English. These inferences are based on the premise that different language speakers have more or less the same capacity of working memory. A more bold inference would be that the mechanisms of cognition are somewhat different in Chinese and English minds, and Chinese may better cope with longer dependency distances than English. The present research method may not sufficiently support this interesting hypothesis, and it requires joint efforts of psycholinguists and neurolinguists in order to provide conclusive evidence.

By carefully comparing the data in Figs. 5–8, we find some fascinating phenomena which deserve more discussion. Data in Figs. 6–8 suggest that the MDD optimization mechanisms are different between English and Chinese. The percentage of adjacent dependencies is about constant at 60 in Chinese, regardless of sentence length, whereas it falls in English as sentence length goes up. On the other hand, the MDD increases in much the same way in both languages (Figs. 6–8). In short, the figures for adjacent dependencies in English fit well into the overall pattern for all dependents: the higher percentage of adjacent dependencies naturally leads to a lower MDD compared to that of Chinese; with the increase of SL, the percentage of adjacent dependencies fall, logically causing the MDD to increase with the increase of SL. But figures for adjacent dependencies in Chinese are not self-explanatory. The overall percentages of adjacent dependencies are lower than those of English, and with the increase of SL, the percentages remain relatively stable. This predicts a higher MDD for Chinese than that for English, which has been confirmed. But why is it that, with the increase of SL, its MDD also increases slightly as shown in Figs. 6–8? A priori expectation would be that its MDD should remain stable with the increase of SL. This indicates different languages have different mechanisms to regulate their MDDs, and some languages (in this case Chinese) tend to have a higher basic MDD, as indicated by “a” in Figs. 6–8.

These differences may suggest the possibility of classifying languages based on dependency distance. To put it differently, can the parameter of “a”, whether in a linear function or a power function, be used as a metric to classify the type of languages? Theoretically, languages which are morphologically richer (i.e. English) may have shorter MDDs than those with a poorer (i.e. Chinese). However, DD has already been found as an unreliable metric for language classification based on the research of DD in over twenty languages (Liu, 2008; Liu and Xu, 2012). Research on modern word order typology reveals that what is closely related to language classification may be the word order that forms grammatical function (Greenberg, 1963; Lehmann, 1973; Song, 2012), called dependency direction under the framework of dependency grammar. Using dependency
direction as a metric for language classification has been verified by Liu (2010) using natural language materials of twenty languages. If dependency direction is a more reliable metric than dependency distance, then, what is its relationship with sentence length?

3.4. The relationship between dependency direction and sentence length

Dependency direction refers to the relative position of two linguistic units (head and dependent) that have a dependency relation in a sentence. If the head follows the dependent, this dependency is called head-final, otherwise, it’s head-initial. In Fig. 1, there are two head-final and two head-initial dependency relations. In this sentence, for example, 50% of the dependency relation is head-final and 50% is head-initial. By the same method, we can calculate the percentage of the two different dependency directions for our 42 sentence sets with different lengths. In actual practice, we need to calculate only the number of one dependency direction, because the sum of the two kinds of dependencies always equals 1. Fig. 9 shows the relationship between head-final dependencies and SL.

Fig. 9 indicates that whether viewing English or Chinese, the percentage of head-final dependency falls gradually with the increase of SL, but there is no statistically significant correlation between the percentage and SL, among which the English linear fitting formula is $y = 0.4859 - 0.0017x$ ($R^2 = 0.0805$, Pearson correlation = $-0.284$, P-Value = 0.213), and Chinese is $y = 0.6675 - 0.0024x$ ($R^2 = 0.1819$, Pearson correlation = $-0.422$, P-Value = 0.057). This indicates that compared with DD, dependency direction is a metric less influenced by the change of SL. Fig. 8 also reveals that there is a significant difference in the number of head-final dependency between English and Chinese, English being 0.467, and Chinese 0.641. Based on this, dependency direction serves as a more suitable metric than dependency distance for language classification and genre judgment. Another observation is that longer sentences tend to have longer MDDs (Figs. 6–8), and longer sentences also have higher percentages of head-initial dependency directions, thus it logically follows that the longer dependencies found in longer sentences tend to be head-initial rather than head-final—i.e., there is a statistical connection between dependency direction and dependency distance. But, it should be noted that this connection is weak or simply a visual connection, because Fig. 9 indicates that SL is not statistically correlated with dependency direction.

Why do we put genre judgment on a par with language classification? From the point of quantitative text analysis, the two bear great similarities. When the text in question belongs to the same language but different genres, we are doing genre (text) judgment. When the text in question belongs to different languages, we are doing language classification or typology. After conducting quantitative studies of five different dependency treebanks in Chinese, Liu et al. (2009b) found that dependency direction is a metric that can better distinguish genres. Using an American National Corpus, Oya (2013) found that the MDDs of different genres in English are different. However, when we analyzed the mean SLs of texts with different genres and their corresponding MDDs extracted from Oya (2013), we found a very high Pearson correlation 0.985. This indicates that sentence length may play a similar role, if not larger, than the dependency direction, since it has been a common consensus to compare genre on the basis of sentence length (Sherman, 1888; Yule, 1939; Kelih et al., 2006).
4. Conclusions and implications

Our data, derived from the parallel treebank of English and Chinese, suggest that within our chosen sentence length, the probability distribution model of dependency distance is not influenced by either the type of language or sentence length; i.e. the dependency distance distributions of 42 sets of sentences with progressive sentence length of English and Chinese follow several distribution models that reflect the distributions of dependency distance have regularities. The main cause for these regularities lies in the fact that human languages have a tendency to minimize the dependency distance, during which the constantly high percentage of adjacent dependencies plays a critical role.

The quantity of adjacent dependencies varies not only with the increase of sentence length but also with the type of language. For Chinese, the percentages of adjacent dependencies fluctuate only within a certain range, instead of showing synchronized change, while for English, the percentages of adjacent dependencies show a falling tendency. The changes and features are subject to the specific type of language, and the change is kept within a stable range from 55% to 67% for English, and from 55% to 64% for Chinese.

Another peculiarity lies in the MDDs. The MDDs of both languages show slight ascending trends with the increase of sentence length, and the MDDs of Chinese are higher than those of English for each sentence set. Our present study provides consistent empirical evidence that MDD of a sentence is related to its sentence length. Since the MDDs of different Chinese sentence lengths are always higher than those of English, and the already-high MDDs are still on a gradual rise with the increase of SL, we propose a conclusion that Chinese taxes working memory more than English. But whether this indicates that Chinese may better cope with longer dependency distances than English still needs further verification from different disciplines. The decrease of adjacent dependencies in conjunction with the increase of MDDs with the increase of SL suggests that English also has long sentences that challenge English users’ working memory, but since the Chinese MDDs are always higher than those of English, this challenge is still less demanding than the one from Chinese long sentences. As for dependency direction, the percentages of head-final dependencies are always higher than those of English for every different sentence set. With the increase of sentence length, the percentages of head-final dependencies show a tendency to decrease in both languages, but the decreases are not significantly correlated with sentence length. It is also found that there is a weak statistical connection between dependency direction and dependency distance—i.e. the longer dependencies found in longer sentences tend to be head-initial rather than head-final. Compared with dependency distance, the dependency direction is a more reliable basis for language and genre classification.

The similarities between two languages that belong to two distinct language families are deserving of our attention and may have far-reaching implications. Some of the explanations to these similarities may lie outside the scope of linguistics, and cause us to seek answers for this universality from a cognitive point of view. Our findings suggest that cognition is a major factor affecting dependency distance, even though language-related factors might affect certain traits of dependency distance. Dependency distance is influenced by cognitive capacity. Moreover, because of the principle of cognitive load minimization, the quantity of adjacent dependencies accounts for more than 50%, causing the MDD to fluctuate within a narrow band despite the increase of SL. These results not only contribute to a better understanding of the languages under study, the relationship between dependency distance and human cognitive mechanisms for language, but also enable the construction of syntactic synergetic subsystems of human languages with greater psychological reality. But, we cannot overlook the peculiarities of the languages since they provide us with information about how our brains may work, especially the differences in coping with the dependency distances of different languages.

The present study demonstrates the possibility of analyzing natural languages to ascertain the relationship between language and cognition, and the human cognitive mechanisms in addition to carrying out psychological experiments. Our in-depth study of dependency distance from the perspective of cognitive capacity may be of interest to linguists as well as scholars in other disciplines such as cognition, communication. The distinctive research method and the interpretation perspective will open up new possibilities for quantitative text analysis, and “advance the research of cross-linguistic complexity and efficiency” (Hawkins, 2014). Using additional larger parallel treebanks in other languages, we may gain a deeper understanding of the universalities and peculiarities of dependency distances across different languages. It is hoped that future research from psycholinguistics and neurolinguistics may provide answers to the questions that cannot be solved purely by corpus and quantitative linguistics.

Acknowledgments

We would like to thank the two anonymous reviewers for their insightful suggestions and comments. This work is partly supported by the National Social Science Foundation of China (Grant No. 11&ZD188).

References

Miller, G., 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97.