Language Sciences 50 (2015) 93–104

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

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

ARTICLE INFO

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

ABSTRACT

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.

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

* Corresponding author. Department of Linguistics, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, CN-310058, China.
E-mail address: lhtzju@gmail.com (H. Liu).

http://dx.doi.org/10.1016/j.langsci.2015.04.002
0388-0001/© 2015 Elsevier Ltd. All rights reserved.

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).[1] 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.