Language Sciences 59 (2017) 135–147

Contents lists available at ScienceDirect

Language Sciences

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

The effects of genre on dependency distance and dependency direction

Yaqin Wang [a], Haitao Liu [a,b,*]

[a] Department of Linguistics, Zhejiang University, Hangzhou, 310058, China
[b] Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, 510420, China

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form 18 September 2016
Accepted 22 September 2016

Keywords:
Dependency distance
Genre
Sentence length
Quantitative linguistics
Working memory

ABSTRACT

Dependency distance, the distance between words linked in a syntactic dependency, has been a key measure of interest in cross-linguistic corpus work, because it is hypothesized to reflect working memory demands during sentence processing. However, previous work has not systematically investigated the effects of text genre on dependency distance in English. We might expect spoken language to have shorter dependencies and informative technical language to have longer dependencies than do fiction and imaginative language. The current study uses quantitative methods to explore the distribution of dependency distance in English across genres controlled for sentence length. Ten genres of English from the British National Corpus provided the data source. The results show that 1) The distributions of dependency distances across all sentence lengths and genres follow the same parametric distributions. 2) Genre affects adjacent dependency rate significantly, but its effect is very small. 3) Sentence length and genre effect dependency distance significantly, but the effect is small. We find shorter dependencies in written-to-be-spoken texts, and longer dependencies in the imaginative genre than those in the informative genres. 4) Genre effects dependency direction significantly, but again the effect is small. Overall the results suggest that the effect of genre on these dependency measures is small, suggesting that dependency distance is primarily determined by universal cognitive factors rather than genre-specific stylistic factors.

1. Introduction

Dependency Grammar describes unequal syntactic relations between two words in a sentence. One of the two words acts as the governor and the other as the dependent (Heringer, 1993; Hudson, 2010; Liu, 2009). The linear distance between the governor and the dependent is the "**dependency distance**" (Heringer et al., 1980), "measured in terms of the number of intervening words" (Hudson, 1995, p. 16). The dependency relation between adjacent words have a dependency distance of 1 and is called **adjacent dependency**. This measure indicates that a word must be kept active in working memory until its meaning has been integrated with the meaning of the word on which it depends (Hudson, 2003; Liu, 2008). Human memory is fleeting and thus the language processing is now-or-never (Christiansen and Chater, 2016). Consequently, the greater the dependency distance, the heavier the memory load the word places on

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

http://dx.doi.org/10.1016/j.langsci.2016.09.006
0388-0001/

the speaker or hearer's mind. The notion of dependency distance thus signals the difficulty level of a given sentence. The greater the dependency distance, the more difficult the syntactic analysis of a sentence is (Gibson, 1998; Hiranuma, 1999; Liu et al., 2009a; Jiang and Liu, 2015).

On the one hand, syntactic complexity has been studied in terms of human working memory by psycholinguists (Jay, 2004; Levy et al., 2013). On the other hand, in the field of cognitive science, syntactic difficulty has been related to the linear distance between governing and dependent words (Gibson, 1998; Grodner and Gibson, 2005; Temperley, 2007; Liu, 2008; Fedorenko et al., 2013; Jiang and Liu, 2015; Lu et al., 2016). In measuring language comprehension difficulty, Yngve (1960) proposes the Depth Hypothesis, which is that there is a relationship between sentence construction and human language memory capacity based on phrase structure. Heringer et al. (1980) then introduced the hypothesis into the framework of dependency grammar. According to the Dependency Locality Theory (Gibson, 1998, 2000), the syntactic complexity of a sentence can be predicted by two factors: the "storage cost" of maintaining the previous words in memory and the "integration cost" of connecting the words to previous words in memory. Other similar theories include Early Immediate Constituents (EIC) (Hawkins, 1994), Minimize Domain (MiD) (Hawkins, 2004), etc. These theories all address the close link between dependency and human language processing.

Dependency distance also indicates the linear order of governor and dependent. Word order has been recognized to be an important means for distinguishing languages with respect to typological features (Greenberg, 1963; Dryer, 1992; Liu, 2010). In the current study, word order is reflected in the **dependency direction**, which indicates whether the dependency relation is governor-final or governor-initial. Hudson (2003) noted that some languages tend to be consistently governor-final (for example, Japanese) and some are governor-initial (for example, Welsh). However, in other languages (such as English and French), some dependents are on the one side of the governor and others are on the other side. Liu (2010) confirmed the validity of measuring dependency directions for language typological classification.

Speakers tend to minimize dependency distance, presumably due to constraints on human working memory and grammar (Ferrer-i-Cancho, 2004, 2016; Buch-Kromann, 2006; Liu, 2008; Gildea and Temperley, 2010). Based on a study of 20 languages, Liu (2008) found that dependency distance minimization occurs across human languages. Based on research on 37 languages, Futrell et al. (2015) also found that this is a universal property of language. (They use the concept of the sum of dependency distances of a sentence.)

Mean dependency distance is an important measure in predicting syntactic difficulty which reflects the cognitive demands of the language concerned (Hudson, 1995). It is computed from sentences and texts, and is affected by language type (Hiranuma, 1999; Eppler, 2005; Liu and Xu, 2012), sentence length (Oya, 2011; Ferrer-i-Cancho and Liu, 2014; Jiang and Liu, 2015), chunking (Lu et al., 2016), the annotation scheme (Liu et al., 2009b), genre (Liu et al., 2009b; Oya, 2013) and grammar (Liu, 2008; Gildea and Temperley, 2010). Therefore, the study of these factors helps us to better understand the relationship between dependency distance and human language and cognition.

On the one hand, the influence that sentence length may have on dependency distances should not be ignored. While investigating the possible disturbing influence of sentence length, Oya (2011) extracted sentences varying in length from less than 10 up to 30 words to compute dependency distances in three different sets of corpora, namely, an English textbook used by Japanese high schools, English essays written by Japanese learners and Parc 700 Dependency Treebank. The study found that average dependency distances of sentences with 10 words and over and less than 20 words are significantly different. Jiang and Liu (2015) chose sentence lengths ranging from 10 to 30 words based on a parallel corpus of Chinese and English. Jiang and Liu found that dependency distance is related to sentence length in those two languages. These studies, nevertheless, do not put emphasis on genres within one language. We still do not know whether the factor of sentence length influences the relationship between genre and mean dependency distance. With regard to the genre factor, on the other hand, Hiranuma's study (1999) demonstrated that more formal texts have longer dependency distances in Japanese than do less formal texts. Liu et al. (2009b) noted that there exist some differences of dependency direction between conversation and written genres of Chinese. The treebank, *cucc*, which uses the conversation genre, demonstrates lower percentages of head-final dependencies than do treebanks composed of written language (Liu et al., 2009b, p. 505, Fig. 7). The question remains unanswered as whether these differences can be seen as a criterion for genre judgment. Oya (2013) examined degree and closeness centrality[1] of dependency trees in a variety of genres of English. The result shows that these distributions of two properties are different for the different genres under fixed sentence length. Oya also mentioned dependency distances of different genres; however, he did not investigate that variable as he had for the other two attributes. The question whether and how genre affects dependency distance, therefore, is still not addressed.

Therefore, providing a larger-scale analysis, the current study makes a quantitative attempt to investigate properties related to dependency relations of various genres under controlled sentence lengths. The result may shed new light on syntactic complexity and features of different genres and also help us to better understand human cognition. We also use a larger corpus than previous studies to increase the robustness of statistical findings.

Ten domains of written English, representing different genres, in the British National Corpus (BNC) (Burnard, 2000) were used as the data source. The following four specific questions are discussed to address the aforementioned issues:

[1] Graph centrality is an index which indicates the difficulty of typed-dependency trees. "*Degree centrality* indicates the extent to which the vertices in a graph are concentrated to one particular vertex… *Closeness centrality* is defined as the reciprocal of the sum of the length of a path from one vertex to another in a graph" (Oya, 2013, p.43–45).