Dependency Distance as a Metric of Language Comprehension Difficulty*

Haitao Liu

Institute of Applied Linguistics Communication University of China
CN-100024 Beijing, China lhtcuc@gmail.com

Linguistic complexity is a measure of the cognitive difficulty of human language processing. The present paper proposes dependency distance, in the framework of dependency grammar, as an insightful metric of complexity. Three hypotheses are formulated: (1) The human language parser prefers linear orders that minimize the average dependency distance of the recognized sentence (2) There is a threshold that the average dependency distance of most sentences or texts of human languages does not exceed (3) Grammar and cognition combine to keep dependency distance within the threshold. Twenty corpora from different languages with dependency syntactic annotation are used to test these hypotheses. The paper reports the average dependency distance in these corpora and analyzes the factors which influence dependency distance. The findings — that average dependency distance has a tendency to be minimized in human language and that there is a threshold of less than 3 words in average dependency distance and grammar plays an important role in constraining distance — support all three hypotheses, although some questions are still open for further research.

Keywords: dependency distance, comprehension difficulty, treebank, cognitive cost


* We thank the anonymous reviewers of JCS for valuable suggestions, Richard Hudson and Probal Dasgupta for insightful discussion and detailed comments, and Anat Ninio, Yue Ming and Li Minglin for helpful comments. We are grateful to Hu Fengguo for generating the random dependency treebank, Joakim Nivre for providing dependency versions of Penn English and Chinese treebanks, Atanas Chanev for sending us the Malt-format of Italian treebank, and to all the providers of treebanks in CoNLL-X’06 and 07.
1. Introduction

Psycholinguistics has provided an empirical basis for the study of language comprehension difficulty (Jay 2004). It is a challenge for formal computational and cognitive linguists to find a metric for measuring this difficulty.

Yngve’s (1960) proposal concerning depth is one attempt to meet this challenge, as his paper examines what we would now call sentence comprehension difficulty. He describes “the maximum number of symbols needed to be stored during the construction of a given sentence” as the depth of that sentence. The paper formulates his Depth Hypothesis as follows: “(a) Although all languages have a grammar based on constituent structure, (b) the sentences actually used in the spoken language have a depth that does not exceed a certain number (c) equal or nearly equal to the span of immediate memory (presently assumed to be 7 ± 2). (d) The grammars of all languages will include methods for restricting regressive constructions so that most sentences will not exceed this depth …” (1960: 452; cited from Yngve 1996: 52). From this quotation, we can extract the following points: Yngve’s hypothesis is based on phrase structure; even if the grammar theoretically permits deeper sentences, in practice the depth of a sentence cannot exceed a certain threshold, which is nearly equal to the capacity of human working memory (Miller 1956; Cowan 2001, 2005); the grammars of all languages have means of keeping most sentences within this threshold. Yngve’s Depth Hypothesis is not unproblematic (Frazier 1985), but the importance of Yngve’s work is that he tried to build a universal metric for language comprehension difficulty. We use the word ‘universal’ because his metric has a close link with cognitive structures which we assume are universal to humanity.

To verify Yngve’s hypothesis, we have to explore, based on as many languages as possible, whether such a threshold exists in language comprehension, and if so, what its magnitude is, what mechanisms prevent sentences from crossing the threshold, and so forth. It is also reasonable to assume that the threshold is a statistical mean or a continuum. Although “for the moment we don’t know what that threshold is” (Hawkins 1994: 13), the search for this threshold is one task in the enterprise of measuring language comprehension difficulty. Further, the threshold, once identified, may also