

Dependency distance as a metric of language comprehension difficulty¹

Haitao Liu

Institute of Applied Linguistics
Communication University of China
CN-100024 Beijing, China
lhuc@163.com

Abstract: 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

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

¹ 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.

(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 facilitate the further task of determining the constant that expresses the capacity of working memory, for "if a constant can be found, it greatly simplifies the theory of human performance." (Cowan 2005: 6)

Miller and Chomsky, in their classic paper (1963), propose a metric of syntactic complexity based on the ratio between non-terminal and terminal nodes of the syntactic tree of a sentence. Frazier (1985) proposes using a local count to replace the global count of Miller and Chomsky as one way to make the metric more sensitive.

Hawkins (1994) assumes a close relation between the grammar and word order. In the context of measuring and predicting syntactic difficulty, he proposes the principle of Early Immediate Constituents (EIC): "The human parser prefers linear orders that maximize the IC-to-non-IC ratios of constituent recognition domain." (p. 77) In later work, Hawkins (2004) updates the EIC and proposes a new principle, "Minimize Domain" (MiD): "The human processor prefers to minimize the connected sequences of linguistic forms and their conventionally associated syntactic and semantic properties in which relations of combination and/or dependency are processed." (p. 31). The new version of the principle proposed by Hawkins postulates relatively direct and obvious links between linear order and human language processing.

This issue about linear order and comprehension difficulty has been investigated in current cognitive science (Gibson 1998, Gibson/Pearlmutter 1998, Gibson 2000, Grodner/Gibson 2005). Gibson (2000) proposes a distance-based theory of linguistic complexity - the Dependency Locality Theory (DLT). There are two key insights in the DLT: "1. Resources are required for two aspects of language comprehension: (1) storage of the structure built thus far and (2) integration of the current word into the structure built thus far. 2. The structural integration complexity depends on the distance or locality between the two elements being integrated." (p.102) The Grodner/Gibson study (2005) reports the results of two reading experiments, which suggest that "the difficulty associated with integrating a new input item is heavily determined by the amount of

lexical material intervening between the input item and the site of its target dependents." (p. 261)

The above-mentioned studies exemplify recent interest in complexity metrics that involve the relation between linear order and syntactic difficulty, and in the role that dependencies between linguistic units plays in the context of a complexity metric. The emphasis on dependency is unsurprising: if human cognitive structure is network-like (Hudson 2007), analysis in terms of a network of syntactic dependencies is an important step towards a conceptual network. It is natural to infer that syntactic dependency structure can map onto cognitive structure and a language network (Liu 2008a) better than the phrase structure assumed in the earlier works. The studies mentioned earlier use the phrase structure approach to syntax. It is certainly possible to construct a framework combining dependency mechanisms and phrase-structural constituency mechanisms, but there is no immediate motivation for attempting this. If our goal is to develop a reasonable metric of linguistic complexity, the point is to consider syntactic representation in the context of parsing strategy issues. Given the prevalence of network conceptualizations in cognitive science, we find it likely that the dependency syntax approach will prove preferable to the phrase structure approach in the long run. We therefore find it appropriate to develop a comprehension difficulty metric that works with tools drawn from the dependency approach to syntactic structure.

The paper begins by appealing to one of the most essential properties of human language – linearity - and accordingly proposes a metric of syntactic difficulty based on treebanks, which can be extended in the context of measuring the complexity of a language. Section 2 presents previous works by other scholars and the method used in the present study. Section 3 introduces the resources used and the major results of this study. More detailed discussions and analysis of the results are presented in Section 4, followed by concluding remarks and suggestions for further research in the last section.

2. Methods

We shall take it for granted in this paper that the syntactic structure of a sentence consists of nothing but dependencies between individual words – an assumption that is widely accepted not only in computational linguistics (Nivre 2006) but also in theoretical linguistics (Hudson 2007). The following are generally accepted as the core properties of a syntactic dependency relation (Tesnière 1959, Hudson 1990, 2007):

1. It is a binary relation between two linguistic units.
2. It is usually asymmetrical, with one of the two units acting as the governor and the other as dependent.
3. It is classified in terms of a range of general grammatical relations, as shown conventionally by a label on top of the arc linking the two units.

Based on these three properties, we can build a syntactic dependency tree or directed dependency graph as the representation of a sentence. In the paper, we use directed acyclic graphs to present dependency structure as in Figure 1. Figure 1 shows a dependency analysis of the sentence *The student has a book*.

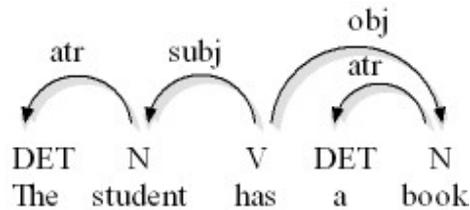


Figure 1. Dependency structure of *The student has a book*

In Figure 1 all the words in a sentence are connected by grammatical relations. For example, the subject and the object depend on the main verb; prepositions (not exemplified in figure 1) depend on the nouns or verbs that they modify; and so on. In each pair of connected words, one is called the *dependent* and the other is called the *governor*. The labeled arc is directed from the *governor* to the *dependent*.

In order for a sentence to be parsed, yielding a dependency structure as in Figure 1, a parsing algorithm must specify the route from the sentence to the structure. There are many parsing algorithms based on dependency grammar (Covington 2001, 2003; Hellwig 2006; Nivre 2006). In the context of modeling human sentence processing, an incremental parsing strategy is preferred. The following is an algorithm adapted from Covington (2003):

This parsing algorithm accepts words one by one, and maintains two lists: *WordList*, which contains all the words seen so far, and *HeadList*, which contains all the words that are not (yet) known to be dependents of other words.

Whenever a word *W* is received from the input list, the parser does the following things:

1. Looks *W* up in the lexicon and creates a node for it.
2. Looks through *HeadList* for words that can be dependents of *W*; attaches them as such, removing them from *HeadList*.
3. Looks through *WordList* for a word of which *W* can be a dependent. If such a word is found, attaches *W* as that word's dependent. Otherwise, adds *W* to *HeadList*.
4. Adds *W* to *WordList*.

Note that steps 2 and 3 can both occur. That is, the current word can acquire dependents in step 2, and then also acquire a head in step 3. When the process is complete, and all the words in the input list have been processed, *HeadList* should have only one element, the main verb.

Take the parsed sentence *I actually live in Beijing* as an example. When the parser is reading the word *live*, *HeadList* includes two words *I* and *actually*; after the parser has processed the word *live*, *HeadList* retains only the verb *live*. We assume that the list *HeadList* serves as the working memory in a human parsing model. In order to ensure successful parsing, the *HeadList* has to handle all words between a dependent and its governor. Thus, this parsing algorithm may be regarded as memory-based. It provides a relatively direct explanation for Gibson's (1998) claim that "the greater the distance between an incoming word and the most local head or dependent to which it attaches, the greater the integration cost."

Current work in psycholinguistics make it reasonable to adopt a working memory approach: "the goal of the parsing model from a working memory view is to provide an analysis of parsing that will explain why some syntactic structures are more difficult to

understand than others...”(Jay 2004: 164).

The linear distance between governor and dependent is defined as “dependency distance²”. Liu/Hudson/Feng (forthcoming) propose a method for measuring the mean dependency distance of a sentence, of a sample of a treebank (a corpus with syntactic annotation) or of a particular dependency type in a treebank. Formally, let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_a and W_b , if W_a is a governor and W_b is its 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 is the case 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 follows the dependent; when a is smaller than b , the DD is a negative number and the governor precedes the dependent. However, in measuring dependency distance the relevant measure is the absolute value of dependency distance.

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

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

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

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

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

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

For instance, a series of dependency distances can be obtained from the sentence in Figure 1 as follows: 1 1 1 2. In other words, the example has three dependencies with $DD = 1$ and one dependency with $DD = 2$. Using Formula (1), the MDD of this sentence is $5/4=1.25$.

Combining the parsing algorithm mentioned above and the formula (1) and (2), we can assume that the greater the MDD of a sentence, the more difficult the sentence. Extending this conclusion to a text (or a language considered as a set of texts), the greater the MDD of a text, the more difficult the text (or language).

It is possible to link the parsing algorithm and the formula for MDD. While a sentence is parsed, we can calculate the dependency distance based on the current item(s) in *Headlist*. For instance, the maximum dependency distance in the sentence *I actually live in Beijing* is between the word *I* and *live*, which is also reflected in *Headlist* if we parse the sentence with the proposed algorithm.

The literature on memory has shown that going beyond a fixed bound often causes problems (Cowan 2001, 2005). Therefore, it seems reasonable to use the maximum DD in a sentence to measure the difficulty of a sentence. Unfortunately, a small bound on the

² The concept first appears in Heringer/Strecker/Wimmer (1980:187), who seem to extract the idea – for which they use the term *Abstand* -- from the Depth Hypothesis of Yngve (1960, 1996). The term ‘dependency distance’ is introduced in Hudson (1995:16) and defined as “the distance between words and their parents, measured in terms of intervening words.”

maximum DD could not be found in these corpora in this study, in other words, it is not feasible to use maximum DD as a stable metrics as language comprehension difficulty. Therefore, we prefer to propose and use a measure which may link dependency distance and the number of open dependencies that need to be stored in the parser's list. In the parsing model proposed, a dependency is open from the moment when it is stored into the *Headlist* of the parser to the point where it is removed from the list³. Hudson (1995) proposes a simple way to measure the load due to open dependencies is to weight each open dependency equally, and to sum the open dependencies after each word (for example, *I actually live in Beijing*): one after *I*, two after *actually*, one after *live*, one after *in*, and zero after *Beijing*. Hudson call the score after a word W_i the *dependency density* at W_i . So, for the exemplified sentence, we can create a series of dependency density 1+2+1+1+0, whose companion of dependency distance is 2+1+0+1+1. Thus, MDD and the mean number of items that are kept in *Headlist* (working memory) during the parsing process are positively correlated. Using MDD, we can measure relative difficulty of a sentence. Sometime it is probably not more precise than using maximum DD to measure absolute difficulty to build a dependency link, but it works with corpus and real text.

Distance is an important property of a dependency relation because of its implications for the cognitive cost of processing the dependency; likewise, the average dependency distance of a text is an important comparative measure and throws light on the cognitive demands of the language concerned relative to other languages. 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. if they are both examples of casual conversation or 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.

At this point, we return to EIC (or MiD) proposed by Hawkins (1994) and DLT due to Gibson (2000) and hypothesize that:

- (1) The human language parser prefers linear orders that minimize the average dependency distance of the recognized sentence or the text.
- (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.

The essential point of these hypotheses is that human languages tend to have a minimized average dependency distance, which is constrained by human cognitive structure (working memory capacity) and grammar. From a complexity viewpoint, the minimization of dependency distance can be regarded as an emergent phenomenon in sufficiently long sentences (Ferrer i Cancho 2008).

To test and verify the hypotheses, we need to work on as large and diverse a linguistic sample as possible, because the hypotheses need to be validated against many languages. The importance of cross-linguistic corpus data is also emphasized in

³ In our model, main governor (verb) of a sentence will remain in the list from this point that be found until the finish of the parsing. It is not difficult to eliminate its influences on the correlation between MDD and the parser's list.

Temperley (2007: 328). The underlying hypothesis we are testing is the claim that this approach is universally valid because it is closely related to human cognitive structure. In addition, research should be based on corpora of real texts, which are a better and more complete reflection of language competence than artificial examples.

Jurafsky (2003:43) finds problems in using corpora to explore psycholinguistic questions. Since a corpus is an instance of language production, how then can the frequencies derived from the corpus be used to draw conclusions regarding language comprehension? However, the aim of this paper is to examine a hypothesis about the use of MDD as a measure of processing complexity. In this way, a treebank is very useful resource to test the hypothesis proposed above, for the reason that it is a collection of real instances of language use. Temperley (2007) shows that DLT (Gibson 2000) is also valid in language production. Therefore, it is reasonable to use a treebank as the resource of our experiment.

There are a few treebank-based studies that investigate the minimization of average dependency distance in a sentence or text.

Ferrer i Cancho (2004), on the basis of a Romanian dependency treebank, hypothesizes and then proves that (a) the average distance of a sentence is minimized and (b) the average distance of a sentence is constrained.

Using a syntactic and two random corpora, Liu (2007) shows that probability distributions of dependency distances can be well captured by the right truncated Zeta distribution and the syntactic treebank has smaller MDD than two random treebanks.

In a study based on a Danish dependency treebank, Buch-Kromann (2006) finds that 44% of all dependents are immediately preceded by their governor, and 88% are separated from their governor by fewer than 5 words. Therefore he suggests that "human grammars exhibit a preference for minimizing the distance between a word and its governor and landing site."(p. 100)

Unlike earlier work (Ferrer i Cancho 2004), our proposals are built around dependency links and are not sentence-based, which we believe is better suited to a universal and network-oriented theory. For example, if one wants to know the proportion of dependency links between adjacent words in a sample (seen as typical of an entire language), one can obtain this only by considering all dependency links in the sample. Thus, despite the availability of formula (1) for the MDD of a sentence, it is formula (2) that is preferred in current research.

In the present study, complexity is measured as average dependency distance in a sentence, which is not similar to Temperley (2007) that measure the complexity using the total length of all dependencies in a sentence. We hope that the current approach may avoid the problem where a longer sentence will certainly have higher complexity.

In this project, by way of exploring the three hypotheses stated above, we try to answer the following questions: Is there an MDD threshold? If so, is its magnitude less than working memory capacity? What determines the syntactic complexity (MDD) of a language (a treebank)? Is it just random linking of words, or is the actual MDD of a corpus smaller than that of a randomly structured version of the same corpus? Moreover, if it is not random, is it determined by users' free choices limited only by sentence length considerations, or are users limited by the language and the genre as well?

The idea that dependency distance, the dependency incremental parsing algorithm

and working memory combine to yield a natural system does follow from the basic assumptions of our approach. However, this system needs to be tested empirically. In this study, we bring a complexity metric based on dependency distance to bear on several examples using formula (1).

In psycholinguistic work, the comprehension of center-embedded structures has been found to be worse than that of right-branching sentences (Miller/Chomsky 1963; Weckerly/Elman 1992; Hudson 1996). Thus, Sentence (1a), with a right-branching structure, is more readily processed than sentence (1b), which involves center-embedding.

(1a) The woman saw the boy that heard the man that left.

(1b) The man the boy the woman saw heard left.

Is the MDD formula (1) adequate for the task of recognizing the degree of complexity of these two sentences? Figure 2 presents their MDDs:.

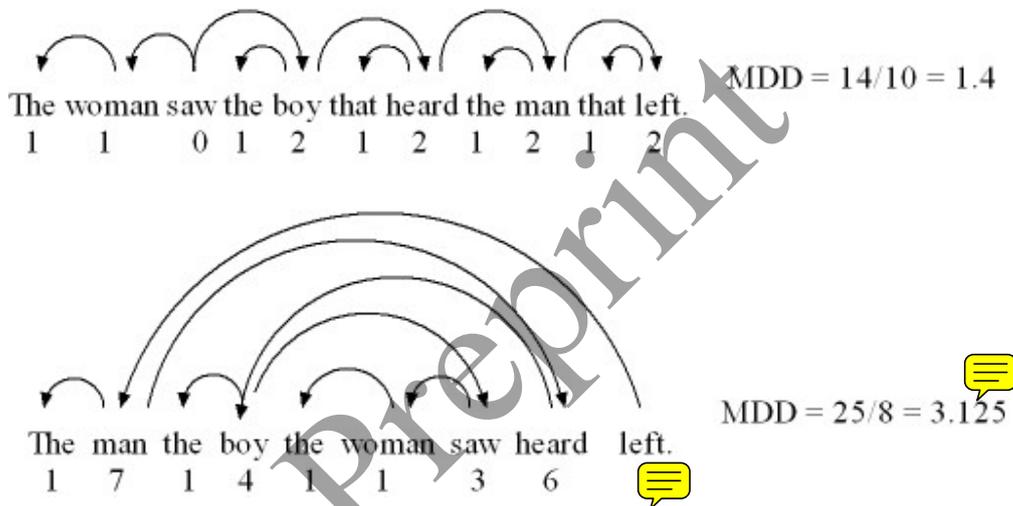


Figure 2. Dependency structures and MDD of right-branching and center-embedding sentences.

The prediction does come out right in this case. The MDD of (1b) is greater than that of (1a), and correspondingly (1b) is more difficult than (1a).

Some types of center-embedded clauses are easier than others. For example, subject-relative sentences (2a) are easier to process than object-relative sentences (2b) (Jay 2003: 165; King and Just 1991).

(2a) The reporter who attacked the senator admitted the error.

(2b) The reporter who the senator attacked admitted the error.

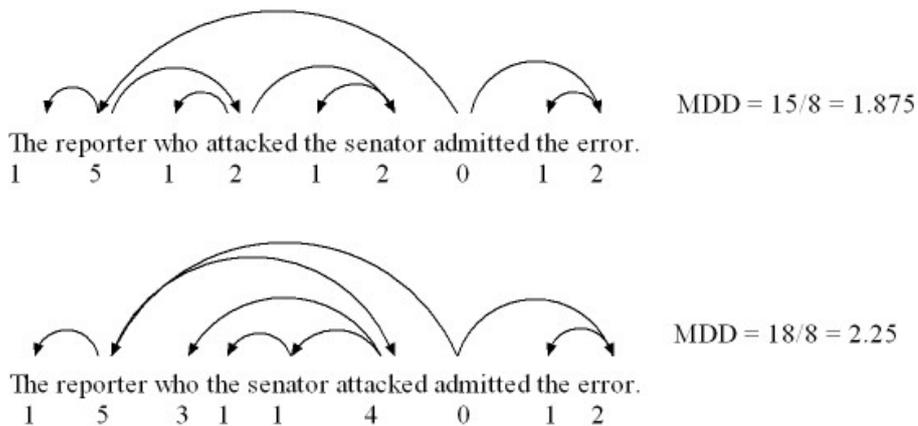


Figure 3. Dependency structures and MDD of subject-relation and object-relative center-embedding sentences.

Again, MDD gives the right results: $1.875 < 2.25$, and (2a) is easier than (2b). These trials suggest that MDD can be used as a complexity metric.

The discussion above shows that MDD can be used to account for processing difficulty in the case of long-distance dependency and center-embedding sentences. However, we also need to consider, in general, how this framework might handle the processing of garden path sentences, and, in particular, how the Late Closure principle works. Consider the garden path example (3):

(3) After the student moved the chair broke.

In sentence (3), the noun phrase “the chair” is initially interpreted as the object of the verb “moved” and is later reanalyzed as the subject of the verb “broke”. The preference for this initial interpretation has been attributed to the Late Closure principle, which is formulated by Frazier (1978) as follows: “When possible, attach incoming lexical items into the clause or phrase currently being processed”. Frazier (1978) suggests that Late Closure has a close link with working memory. “It is a well-attested fact about human memory that the more structured the material to be remembered, the less burden the material will place on immediate memory. Hence, by allowing incoming material to be structured immediately, Late Closure has the effect of reducing the parser’s memory load” (Frazier, 1979, 39). In our terms, the Late Closure principle is preferred by the parser because it tends to minimize the average dependency distance of the sentence being processed. Figure 4 shows the dependency analysis and computes the MDD for (3).

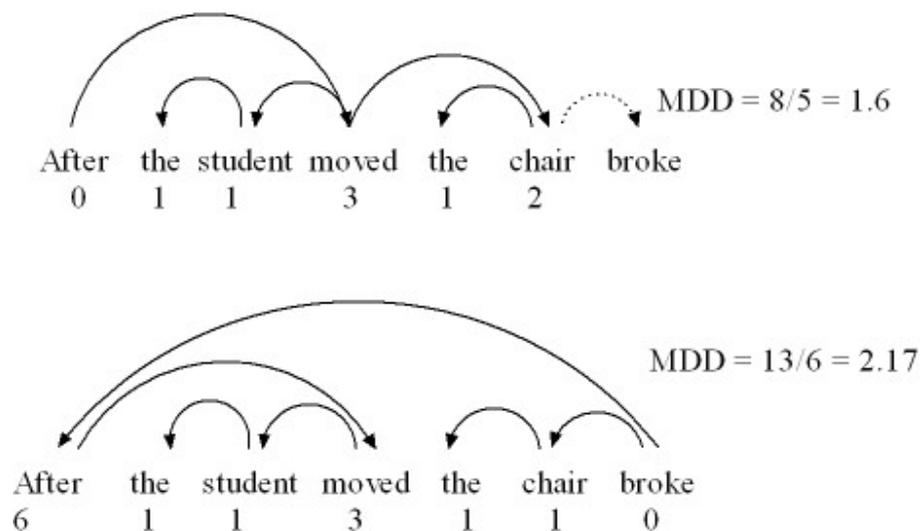


Figure 4. Dependency structures and MDD of garden-path sentence.

Before encountering the word “broke”, the MDD of the parse is 1.6. When the word “broke” is accepted as input, the parser has to break the link between “moved” and “chair”, and reanalyze “chair” as the subject of “broke” and “after” as an adverbial of “broke”; these moves raise the magnitude of the MDD and increase the difficulty of the sentence. This example shows that MDD can adequately detect and measure the difficulty of a garden-path sentence⁴.

In the next section, we present the results of our statistical analysis of dependency distance for 20 languages.

3. Measuring Dependency Distance

We used formula (2) to measure the average dependency distance for 20 languages. These 20 languages are⁵: Chinese (chi), Japanese (jpn), German (ger), Czech (cze), Danish (dan), Swedish (swe), Dutch (dut), Arabic (ara), Turkish (tur), Spanish (spa), Portuguese (por), Bulgarian (bul), Slovenian (slv), Italian (ita), English (eng), Romanian (rum), Basque (eus), Catalan (cat), Greek (ell), Hungarian (hun).

Most of the treebanks used in our project are drawn from the training sets of the CoNLL-X Shared Task on Multilingual Dependency Parsing⁶ (2006, 2007). These treebanks have different annotation schemes, but we use their dependency formats converted by CoNLL-X 2006 and 2007 organizers. Detailed information on all treebanks used is provided in the “Resources” section of this paper.

For all 20 languages, we built a Pareto chart of dependency distance in order to

⁴ It is more difficult to explain counterexamples to the Late Closure principle in the proposed framework (Cuetos and Mitchell 1988).

⁵ Language codes are following ISO 639-2: Codes for the Representation of Names of Languages.

http://www.loc.gov/standards/iso639-2/php/code_list.php (2008-7-6)

⁶ <http://nextens.uvt.nl/~conll/> Tenth Conference on Computational Natural Language Learning - New York City, June 8-9, 2006. <http://depparse.uvt.nl/depparse-wiki/SharedTaskWebsite> for 2007 (2008-7-6). In other words, the project is a by-product that emerged from the CoNLL-X tasks done by this author.

explore the relation between cumulated value and dependency distance. Average dependency distance (MDD) and the distribution of dependency links are also calculated by means of formula (2).

To ascertain the relations between dependency distance and other factors, we also constructed two random dependency treebanks. Ideally, we could have chosen to generate a language with random lexicon and sentences, but it is difficult or even impossible to analyze such a language syntactically. Randomly assigning governors over all words in a dependency Treebank yields a satisfactory random dependency treebank as a sample of a hypothetical random language. In this way, we can calculate MDD of some random treebanks using formula (2). Liu and Hu (2008) provides a detailed formal description and algorithm for generating two random languages.

We use two methods to generate two random treebanks. In the first random treebank (RL1), within each sentence we select one word as the root, and then for every other word we randomly select another word in the same sentence as its governor, disregarding syntax and meaning. Figure 5 shows a random analysis produced on this basis for the sentence in Figure 1.

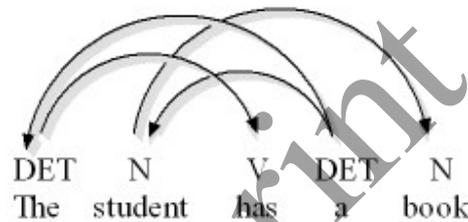


Figure 5. A random analysis of *The student has a book* with crossing arcs

In the second random treebank (RL2), while governors are assigned randomly, we always make sure that the dependency tree (graph) that results is a projective and connected tree, i.e., no crossing arcs are allowed in the graph. This property of a graph is also called *projectivity* and was first discussed in Lecerf (1960) and Hays (1964). Figure 6 gives an example of RL2.

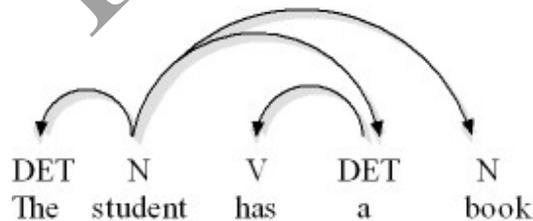


Figure 6. A random analysis of *The student has a book* without crossing arcs

Thus, three dependency graphs can be constructed for the sentence *The student has a book*. The first, shown in Figure 1, is syntactically well-formed; RL1, shown in Figure 5, has the lowest syntactic well-formedness, and RL2, in Figure 6, exceeds RL1 in syntactic well-formedness.

Table 1 lists information relating to the 20 languages investigated. In the table, *size* is the number of dependencies in the sample for this language; *m_{sl}* is the mean sentence length; *l_{dd%}* is the percentage of dependencies between adjacent words; *m_{dd}* is the mean dependency distance of the syntactic treebank calculated using formula (2) (NL); *m_{ddr1}*

is the mean dependency distance of random language 1 (RL1); *mddr2* is the mean dependency distance of random language 2 (RL2); *genre* represents the text genre of the treebank, with *mixed* indicating that the treebank includes different text genres; *type* shows the original annotation scheme of the treebank with D for dependency structure, C for constituent structure, and CF for a mixed structure with constituent and grammatical functions.

Table 1. Statistics for dependency distances and other contrasts in corpora from 20 languages

| language | size | msl | ladd% | mdd | mddr1 | mddr2 | genre | type |
|------------------|--------|------|-------|-------|--------|-------|--------|------|
| Romanian (rum) | 32108 | 8.9 | 67.5 | 1.798 | 5.036 | 2.84 | news | D |
| Japanese (jpn) | 108977 | 7.9 | 80.2 | 1.805 | 6.726 | 3.212 | dialog | CF |
| Danish (dan) | 38120 | 15.9 | 62.3 | 2.136 | 10.500 | 3.948 | mixed | D |
| Italian (ita) | 33690 | 24.4 | 72.4 | 2.19 | 12.251 | 4.659 | mixed | D |
| Bulgarian (bul) | 147071 | 12.5 | 57.5 | 2.245 | 5.542 | 2.876 | mixed | C |
| Turkish (tur) | 38706 | 9.3 | 64.2 | 2.322 | 3.356 | 3.071 | mixed | D |
| Swedish (swe) | 160273 | 15.5 | 51.4 | 2.382 | 8.243 | 3.754 | mixed | D |
| Czech (cze) | 992651 | 14.8 | 53 | 2.441 | 7.953 | 3.671 | news | D |
| Greek (ell) | 55953 | 24.2 | 51.9 | 2.449 | 11.395 | 4.347 | mixed | D |
| Portuguese (por) | 168522 | 19.6 | 55.3 | 2.506 | 12.824 | 4.626 | news | CF |
| Dutch (dut) | 479677 | 12.6 | 51.4 | 2.524 | 6.451 | 3.187 | mixed | CF |
| English (eng) | 152882 | 20.9 | 51.3 | 2.543 | 9.915 | 4.093 | news | C |
| Basque (eus) | 47498 | 15.8 | 55.5 | 2.552 | 7.130 | 3.384 | mixed | C |
| Slovenian (slv) | 22380 | 15.5 | 49.8 | 2.59 | 9.501 | 3.904 | novel | D |
| Arabic (ara) | 50097 | 35.3 | 64.5 | 2.595 | 18.474 | 5.479 | news | D |
| Catalan (cat) | 365530 | 28.8 | 55.2 | 2.645 | 13.126 | 4.683 | mixed | C |
| Spanish (spa) | 75571 | 24 | 55.2 | 2.665 | 13.050 | 4.758 | mixed | CF |
| German (ger) | 564549 | 15.4 | 44.4 | 3.353 | 8.935 | 3.793 | news | CF |
| Hungarian (hun) | 105430 | 21.8 | 46.7 | 3.446 | 11.311 | 4.356 | news | C |
| Chinese (chi) | 412191 | 22.9 | 47.9 | 3.662 | 15.851 | 5.044 | news | C |

The analysis of the data shown in Table 5 is presented and discussed in the following section.

4. Discussion

In section 2, we proposed using the (average) dependency distance of a sentence or a text as the metric for its comprehension complexity or difficulty. Section 3 provided the data drawn from 20 languages. This section will discuss the following questions: is there an MDD threshold? If so, is its magnitude less than working memory capacity? What determines the syntactic complexity (MDD) of a language (a treebank)? Is it just random linking of words, or is the actual MDD of a corpus smaller than that of a randomly structured version of the same corpus? Moreover, if it is not random, is it determined by

users' free choices limited only by sentence length considerations, or are users limited by the language and the genre as well? How do grammar, annotation scheme and language typology influence MDD of a language?

One of the most crucial issues is: do the empirical data in section 3 support the existence of an MDD threshold? Considering such distance to be closely related to the capacity of human cognition, particularly to that of working memory (Miller 1956; Cowan 2001, 2005), we assume that the average dependency distance within human languages should remain below a threshold, which we expect to be smaller than 4 (Cowan 2001, 2005). On the basis of Table 1, we can draw a chart in terms of MDD for 20 languages.

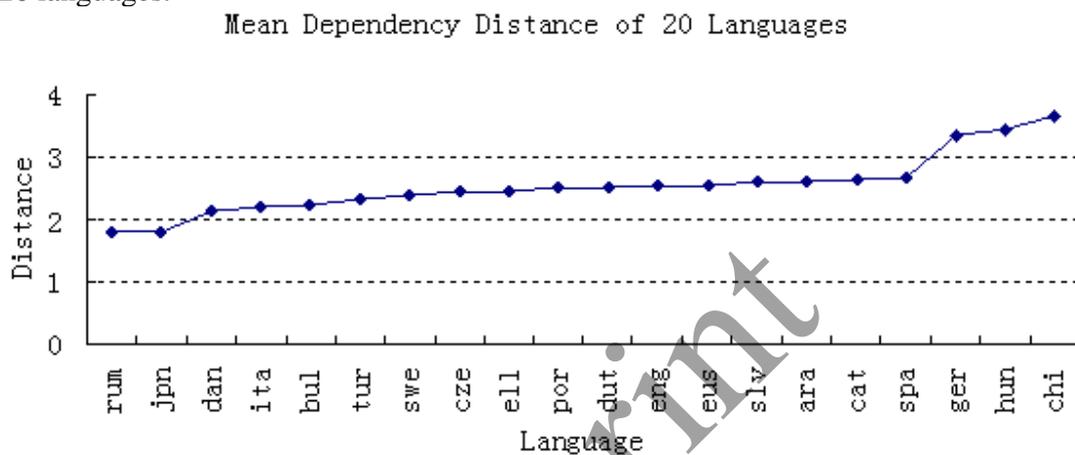


Figure 7. MDD of 20 languages.

Figure 7 clearly shows that for 20 languages MDD only varies over a small range, in spite of the fact that the language samples investigated exemplify different typological features and text genres. Among these languages, Chinese has the greatest MDD. This finding is consistent with the results mentioned in Liu/Hudson/Feng (forthcoming), a study based on a distinct sample.

Given Figure 7, it seems reasonable to infer that human languages have a minimized MDD threshold and it is within working memory capacity (Cowan 2001, 2005). If this is considered as a property of human language, a random language should have a greater MDD than natural language. Therefore, we make a comparison of MDD between human language and two random languages in Figure 8.

Figure 8 shows that for each language the two random analyses have much greater MDD than the syntactic (NL) equivalent they are based on. Of the two random languages, RL2 has a smaller MDD than RL1. It is worthy to investigate why RL2 has smaller MDD than RL1 and what role a no-crossing arc plays in this process.

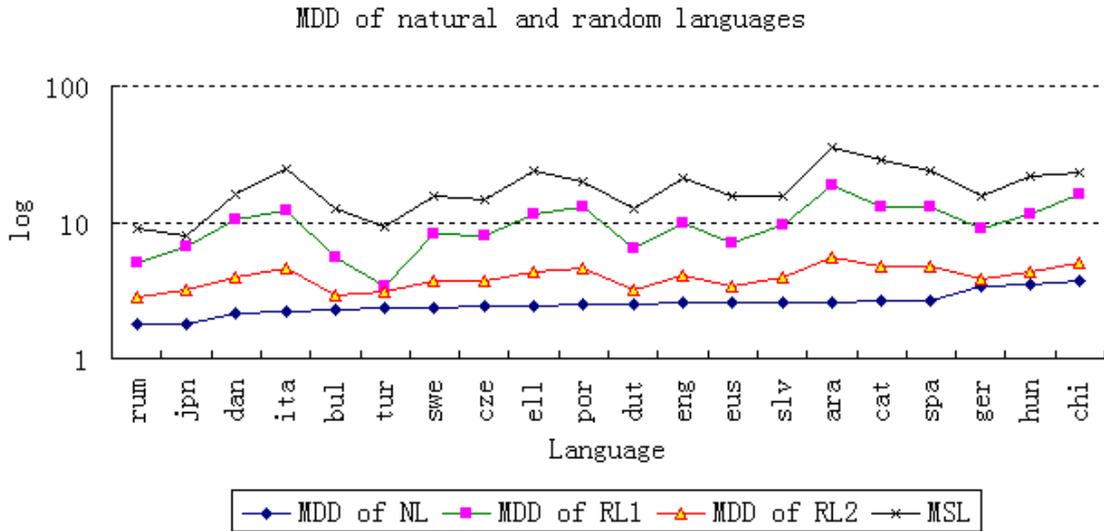


Figure 8. MDD of naturally and randomly structured samples. (Y-axis in logarithm)

The fact that the MDD for RL2 is smaller in our experiment than that for RL1 supports Ferrer i Cancho’s (2007) argument that the uncommonness of crossings in the dependency trees could be a side-effect of minimizing the Euclidean distance between syntactically related words. This is confirmed by our experiment, because RL2 consistently has a smaller MDD than RL1. However, the fact that the variation of MDD for RL2 does not seriously exceed the variation figures for NL (2.8 to 5.0 as compared to 1.8 to 3.6) also verifies the usefulness of a no-crossing-arc approach to MDD reduction, so one could equally reverse Ferrer i Cancho’s argument by considering a reduced distance as a consequence of avoiding crossing dependencies.

If the actual MDDs of NL and RL2 are smaller than that of a randomly structured version of the same corpus, it might indicate that the links of words are not random. If this is the case, then are they determined by users’ free choices limited only by sentence length considerations? To answer this question, MSL of 20 languages are added into Figure 8. Table 2 presents correlation coefficients between MSL and MDD of natural languages, RL1 and RL2.

Table 2. Correlation coefficients between MSL and MDD of natural languages, RL1 and RL2.

| | NL | RL1 | RL2 |
|-------------------------|-------|-------|-------|
| Correlation coefficient | 0.408 | 0.914 | 0.924 |
| P-Value | 0.074 | 0.000 | 0.000 |

Not surprisingly, MSL is very closely related to the MDD of a random analysis, as can be seen in Table 2; but in spite of the low correlation in this table for the natural analyses, we believe (following Ferrer i Cancho 2004) that MSL is in fact correlated with

MDD, and especially so when MSL is short⁷, as we explain below. However, Figure 8 shows that the MDD for NL is numerically much less highly correlated with MSL than the two random languages. This may suggest that the systematic patterns of grammar and other factors also play additional roles in reducing average dependency distance. In other words, Figure 8 provides a functional explanation for syntactic word-order restrictions: one of their (many) benefits is to reduce the MDD of a sentence. It seems that no-crossing-arc and grammar work together to allow us to use long sentences without greatly increasing the MDD within an acceptable range.

If sentence length is not the only factor influencing MDD in NL as shown in Figure 8 and Table 2, could linguistic typology contribute something to this game? Comparing the curve of MSL with that of MDD for NL in Figure 8, we can see that the correlation varies greatly from language to language; for example, Arabic combines a very high MSL with a fairly low MDD, while German shows the opposite pairing. An important reason for studying the correlation between MDD and sentence length is to make it possible to inquire how many of the cross-language differences can be explained in terms of sentence length. In order to explain some of the exceptions seen in Figure 8 and to explore further the relation between sentence length and MDD, we classified languages into two types: low MDD and high MDD, and we investigated the correlations for each type separately. The languages in the low MDD group are Danish, Italian, Portuguese, English, Arabic, Greek, Catalan and Spanish. They are included in this group because their MSLs are less closely related with their MDD than we find for languages in the other group. In contrast, the high MDD group (Romanian, Japanese, Bulgarian, Turkish, Swedish, Czech, Dutch, Slovenian, German, Hungarian, Basque and Chinese) exhibit close relations between MSL and MDD. Table 4 shows that in the high MDD group, some languages even have similar correlation coefficients to their random equivalents.

Table 3. Correlations between MSL and MDD of natural languages, RL1 and RL2 in the low MDD group of languages.

| | NL | RL1 | RL2 |
|-------------------------|-------|-------|-------|
| Correlation coefficient | 0.537 | 0.853 | 0.876 |
| P-Value | 0.170 | 0.007 | 0.004 |

Table 4. Correlation coefficients between MSL and MDD of natural languages, RL1 and RL2 in high MDD group.

| | NL | RL1 | RL2 |
|-------------------------|-------|-------|-------|
| Correlation coefficient | 0.894 | 0.891 | 0.905 |
| P-Value | 0.000 | 0.000 | 0.000 |

Tables 3 and 4 also show that MSL is positively correlated with MDD for all languages; but, for some of the languages, MSL is a more critical factor impacting on MDD. Considering the typological diversity of the languages discussed, the relationship between MSL and MDD may be dependent on the typology of a language; for example,

⁷ Japanese has a shorter MSL due to its dialog genre. Excluding all punctuations, Romanian has also a very short MSL.

most Romance languages are in the low MDD group. However, we need more treebanks in the same language family to explore how linguistic typology can influence the relation between MSL and MDD.

The MDD's comparisons among NL, RL1 and RL2 reveal that, with the exception of MSL and no-crossing arc, grammar also plays an important role to reduce MDD of a sentence (language). Although it is not the primary task of this paper to scrutinize the interesting problem, a few examples perhaps are appropriate.

Based on the Wall Street Journal portion of the Penn Treebank, Temperley (2008) investigates the relation between grammar and dependency distance and introduces three principles of dependency-length minimization in grammar: (1) dependencies should be consistently right-branching or left-branching; (2) shorter dependent phrases should be closer to the head; (3) some "opposite-branching" of one-word phrases is desirable. His computational experiments show that all three principles contribute significantly to dependency-length reduction. His findings are very useful to understand how grammar rules work together to reduce MDD in a sentence. It is our aim in future work to integrate his findings into our model.

An alternative, in which to explore the relation between MDD and language has to do with what MDD one would expect a language to exhibit given its grammatical profile. It would be interesting to see what MDD reduction strategies are employed in the languages in question. For example, Japanese is a governor-final language, which might be expected to have a higher MDD because more dependents are separated from the governor than in a language like English where the dependents tend to occur on either side of the governor. Therefore, it is puzzling to us that Japanese comes out as a language with a high number of adjacent dependencies and lower MDD compared to English (Figure 7 and 10). In our sample, it is not easy to answer the question, because we are using corpora with different genres and annotation schemes. A possible answer for this question is found in Hiranuma (1999), which measures the MDD of conversational English and Japanese. Hiranuma reports that conversational Japanese in fact has a similar MDD to English and suggests, as explanation, that this is because Japanese has a smaller number of dependents than English because it allows more dependents to be omitted.

Among the 20 languages we have investigated, Chinese has the greatest MDD. This may be explained from a grammatical standpoint. For example, whereas prepositional phrases follow the modified noun in English, in Chinese they precede it, which means that the preposition's complement inevitably separates it from the modified noun; moreover, some syntactic functions are realized by the inflection in English, but in Chinese the same functions are often handled by function words which may separate the modified word from its head. Consider the English sentence *I saw the film*, which has the Chinese translation "我看过这部电影"(wo kan guo zhe bu dianying). The DD of the relation between verb and object 'saw-film' is 2, but in Chinese the DD of '看-电影' is 4. Figure 9 diagrammatically shows such difference between these two languages.

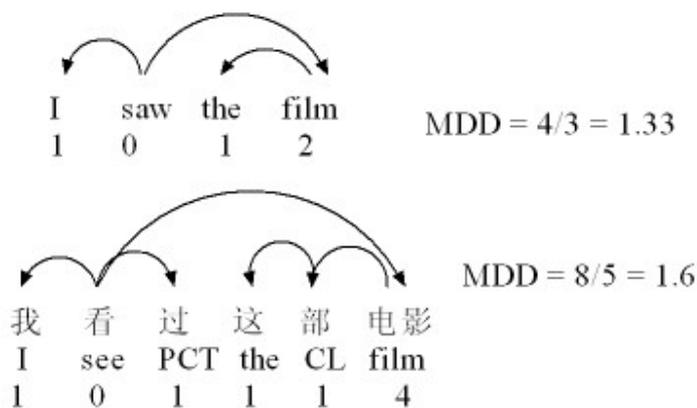


Figure 9. the dependency structures of *I saw the film* and its Chinese translation

The above examples show the need to explore how grammar influences the MDD of a language. Such studies are useful in the context of ascertaining the role of syntax in minimizing MDD. This is an important task for further research.

Another property of human language, that may reduce MDD, is that dependency link tends to be built between adjacent words. It is difficult to classify that this property is a grammatical or typological, but it works.

Dependencies in dependency grammar need not link adjacent words; for example, Schubert defines the notion “dependency” as “directed co-occurrence” (1987: 29), using *co-occurrence* to include not only *adjacent* dependency links, but also *non-adjacent* ones. Consequently, the number of dependencies between adjacent words can influence the MDD of a language. If a language includes many adjacent dependencies, it will have a lower MDD.

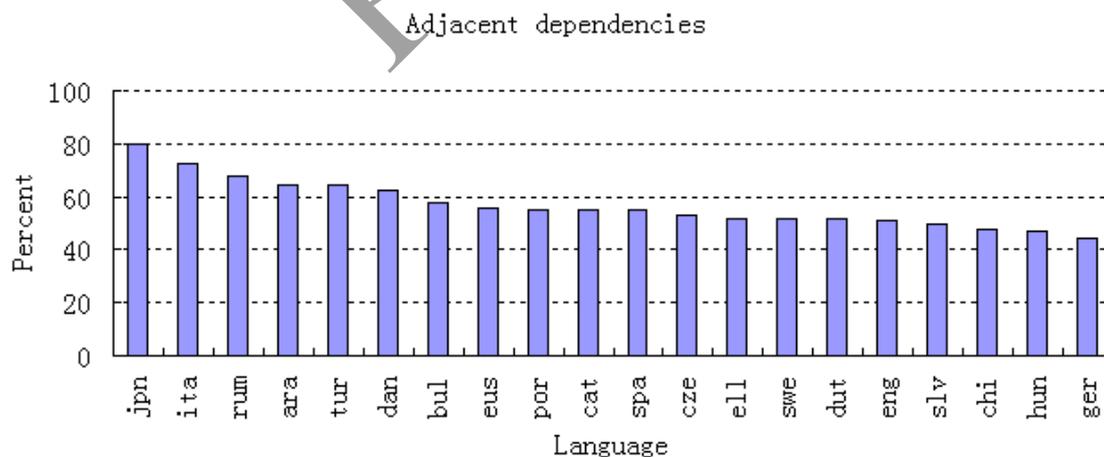


Figure 10. Distribution of adjacent dependencies in 20 languages

Figure 10 reveals, in all 20 languages, that almost 50% of links are adjacent. Some

languages have an even higher percentage of adjacent links, but although the figures may be explicable in terms of grammatical typology in Japanese, Arabic and Turkish, we cannot explain why Italian and Romanian also have higher percentages than other languages in the same language family. The statistics on adjacent and non-adjacent links also demonstrate that it is not reasonable to build a complex syntactic network of human languages exclusively based on adjacent co-occurrences as suggested in Ke and Yao (2008) and the project **WIENER**⁸, because we can not find a corresponding syntactic theory that gives such a salient role exclusively to adjacent co-occurrences. Figure 10 also provides evidence that strongly suggests that this kind of linguistic theory may not accurately describe natural language.

Preference for building dependency between adjacent words or the minimized distribution of dependency distance can also be seen in a time-series plot of all the dependencies in a single treebank. Figure 11 shows such distributions for the Arabic, Slovene and Turkish treebanks.

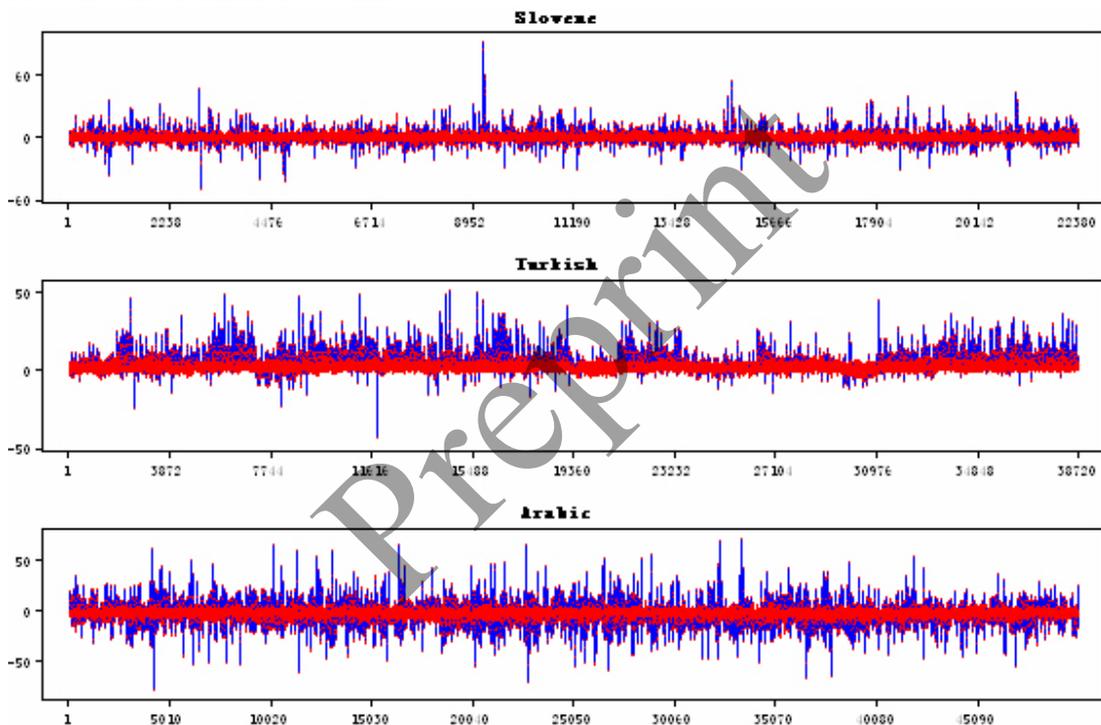


Figure 11. Time Series Plot of dependencies in Arabic, Slovene and Turkish treebanks. X-axis is ordered number of dependencies in the treebank, Y-axis is dependency distance with direction.

Figure 11 reveals that the distribution of dependency distance in a treebank is centered around zero, i.e., it is minimized. The distribution is not balanced above and below zero, and the bias seems to depend on the structure of each language. In other words, Figure 11 not only helps us to demonstrate the minimization of dependency distance in a sentence, a text or a language, but also allows us to observe the effect of a language being a governor-final, governor-first or mixed language. For instance, Figure 11 shows that Turkish is a typical governor-final language, though with a small fringe

⁸ "Word Interactions: Exploring NETWORK Robustness" <http://complex.upf.es/~bernat/WIENER> (2008-7-6)

(below the zero axis) of governor-first dependencies.

Based on the above discussions, we can conclude that MSL, no-crossing arc, grammar and dependency link between adjacent words contribute together to influence and reduce MDD of a sentence (language). However, dependency grammar is not a united syntactic theory, so we have to consider the issue of whether the annotation scheme is likely to have some influence on the MDD of these languages. The point is that what counts as a dependency relation in one treebank does not automatically carry over to other treebanks. For instance, in a noun phrase is it the determiner that is dependent on the noun or the noun that is dependent on the determiner? Both analyses are available in different dependency grammars (Hudson 2004). In the sample of the present study, the determiner is shown as the governor of a noun only in Danish and Italian, but in the other languages the noun is chosen as the governor of the noun phrase. In fact, the question of how to decide the governor in a noun phrase (including the determiner and the noun) is one of the few major disagreements in dependency grammar. This uncertainty, even though it is rare in dependency grammar, will influence the MDD of the related treebanks and make cross-linguistic comparisons unreliable. Thus, we have to inquire whether this influence is likely to affect the conclusion based on the approach proposed here. As the ‘style’ column in Table 1 shows, in our collection, nine treebanks had been constructed in genuine dependency style, five were originally in styles coding function and constituent structure (or constituency), and six were in constituency styles. However, all non-native dependency treebanks used in this study have been converted into dependency treebanks. In other words, all treebanks have been brought into alignment with the three core properties of dependency relations mentioned in section 2. It might be preferable to base conclusions only on treebanks using the same annotation schemes. There are four treebanks using the Prague Dependency Treebank annotation scheme: Czech, Arabic, Greek and Slovene. There are two treebanks using the Penn Treebank annotation scheme⁹: English and Chinese. If the annotation scheme has an important influence on the features above mentioned, then Czech, Arabic, Greek and Slovene would have similar characteristics, in contrast with English and Chinese. However, the tables and figures tell us that the relevant languages are not similar to each other.

Another way to investigate the relation between annotation schemes and the proposed method is to use several treebanks with different annotation schemes and text genres as a resource for the study of one language. Liu (2008b) uses five treebanks for calculating MDD and other syntactic features of Chinese based on the same methods as this paper, and shows that, although the annotation scheme does have some impact on MDD, nonetheless the conclusions from the five treebanks are similar.

Based on these discussions and experiments, we argue that the annotation scheme does influence MDD, but that this effect is not strong enough to affect the conclusion seriously. In a study using corpora, a few local properties of a sentence do not suffice to change the global features of a language. Therefore, the MDD of a language as computed by the methods proposed here more closely depends on linguistic structure and universality of human cognitive structure than on the specific annotation schemes used in the treebank.

⁹ For English and Chinese treebanks, we use the dependency format converted by a similar algorithm.

5. Concluding remarks

We propose dependency distance as the metric of linguistic complexity or language comprehension difficulty. We have investigated the issue on the basis of treebanks (corpora with dependency annotation) for 20 languages. We chose a corpus-based method because if language comprehension difficulty is closely linked with human cognitive structure or working memory capacity, the metric should be universally applicable to, and therefore should be tested against, real texts of human languages.

In order to arrive at an objective approach to the measurement of language comprehension difficulty, we have considered the issue from both a static and a dynamic viewpoint. The static method deals with the representation of syntactic structure, and the dynamic method addresses the parsing algorithm or the cost of converting a linear sentence into syntactic structure.

The static and dynamic aspects are not independent. In section 2 we claimed that if the metric is (dependency) distance-based, then dependency structure and the incremental parsing algorithm would be more suitable candidates. Building on the works of Hawkins (1994) and Gibson (2000), we hypothesized as follows:

- (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.

To test these hypotheses, we have measured the average dependency distance (MDD) and other features of 20 languages. All MDDs fall within the range of 1.798-3.662 – well below the range of MDDs for randomly analysed corpora. The minimized distribution of dependency distance can also be seen in time series plots of all dependencies in the languages explored. Our data reveal that the human language parser prefers to minimize the average dependency distance of a sentence, a text or a language. Among the 20 languages, Chinese has the greatest MDD of 3.662, which means that even in this language we have to memorize on the average less than three words for processing Chinese sentences. For other languages investigated, the value is smaller.

Assuming that the MDD and the mean number of items that are kept in *Headlist* (memory) during the parsing are positively correlated, we can associate MDD with working memory capacity, which has a value around 7 ± 2 (Miller 1956) or 4 (Cowan 2001, 2005). Note that, if we want to link the MDD of a text with working memory, then it is a reasonable procedure to convert a text into a set of sentences. While the greatest MDD (Chinese, 3.66) in our sample is smaller than Cowan's number 4, such a close approximation may be more than a coincidence.

Comparing the MDDs among the treebanks of one natural and two randomized versions for each language, we can observe that grammar, formal constraints (projectivity), and cognition (working memory capacity) work together to keep the mean DD of a language within the threshold.

Our research also adds new concepts to the long list of conceptualizations regarding language economy and minimization (Roelcke 2002: 28). However, there are still many

open questions. For example, if we end up with a concrete number for the mean DD of each language, how can we be sure that this is less than what we might have otherwise expected? What kind of grammatical constraints serve to reduce MDD of a sentence (or language)? Are these grammatical constraints language-dependent or language-independent? How can we reduce the interference of annotation schemes with the proposed method in this study? How can we fine-tune the results by using more normalized corpora? How to evaluate and test our model by direct empirical tests of language production or comprehension under controlled circumstances? How to more closely link our model to general cognitive theory? If MDD has a close link with working memory capacity, why does Chinese has a greater MDD than other languages? If grammar is a means to reduce MDD, why does Chinese grammar not to do the same as other languages?

Answers to these questions in future researches would help to get more interesting findings with more linguistic universality, which in turn could bring us gradually nearer to an understanding of how grammatical and cognitive constraints serve to reduce the MDD of a sentence or language.

Resources

Prague Dependency Treebank (PDT, Czech); Prague Arabic Dependency Treebank (PADT); Slovene Dependency Treebank (SDT); Danish Dependency Treebank (DDT); Swedish Talbanken05; Turkish Metu-Sabancı treebank; German TIGER treebank; Japanese Verbmobil treebank; The Floresta sintá(c)tica (Portuguese); Dutch Alpino treebank; Spanish Cast3LB; Bulgarian BulTreeBank; Romanian dependency Treebank¹⁰; English Penn Treebank, Penn Chinese treebank¹¹; Turin University Treebank (TUT)¹²; Basque Treebank; CESS-Cat Catalan treebank¹³; Szeged Treebank (SzTB)¹⁴; Greek Dependency Treebank (GDT). These treebanks are described in the following documents.

Aduriz, I., Aranzabe, M., Arriola, J., Atutxa, A., Diaz de Ilarraza, A., Garmendia, A., Oronoz, M. (2003) Construction of a Basque Dependency Treebank. TLT 2003. Second Workshop on Treebanks and Linguistic Theories, Vaxjo, Sweden, November 14-15.

Afonso, S., E. Bick, R. Haber, and D. Santos. (2002) "Floresta sintá(c)tica": a treebank for Portuguese. Proc. of LREC-2002, 1698–1703.

Atalay, N. B., K. Oflazer, and B. Say. (2003) The annotation process in the Turkish treebank. Proc. of LINC-2003.

Bosco, C., V. Lombardo, D. Vassallo, and L. Lesmo. (2000) Building a treebank for Italian: a data-driven annotation schema. Proc. 2nd International Conference on Language Resources and Evaluation LREC 2000, pages 99-105, Athens, 2000.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith. (2002) The TIGER treebank. Proc. of TLT-2002.

Csendes, D., Csirik, J., Gyimóthy, T., Kocsor, A. (2005) The Szeged Treebank. V. Matousek et al. (Eds.): Proceedings of the 8th International Conference on Text, Speech

¹⁰ <http://phobos.cs.unibuc.ro/roric/texts/indexen.html> (2008-7-6)

¹¹ <http://www.cis.upenn.edu/~chinese/ctb.html> (2008-7-6)

¹² <http://www.di.unito.it/~tutreeb/> (2008-7-6)

¹³ <http://www.lsi.upc.edu/~mbertran/cess-ece> (2008-7-6)

¹⁴ <http://www.inf.u-szeged.hu/hlt> (2008-7-6)

and Dialogue, TSD 2005, LNAI 3658, pp. 123-131, Springer Verlag, 2005.

Civit Torruella, M. and Ma A. Martí Antonn. (2002) Design principles for a Spanish treebank. Proc. of TLT-2002.

Dzeroski, S., T. Erjavec, N. Ledinek, P. Pajas, Z. Zabokrtsky, and A. Zele. (2006) Towards a Slovene dependency treebank. Proc. of LREC-2006.

Hajic, J., O. Smrz, P. Zemanek, J. Snidauf, and E. Beska. (2004) Prague Arabic dependency treebank: Development in data and tools. Proc. of NEMLAR-2004, 110–117.

Kawata, Y. and J. Bartels. (2000) Stylebook for the Japanese treebank in VERBMOBIL. Verbmobil-Report 240, Seminar für Sprachwissenschaft, Universität Tübingen.

Nilsson, J., J. Hall, and J. Nivre. (2005) MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. Proc. of the NODALIDA Special Session on Treebanks.

Oflazer, K., B. Say, D. Zeynep Hakkani-Tur, and G. Tur. (2003) Building a Turkish treebank. Abeillé A. (ed.) Treebank: Building and using Parsed Corpora, Dordrecht: Kluwer. chapter 15.

Prokopidis, P., Elina Desypri, Maria Koutsombogera, Haris Papageorgiou, and Stelios Piperidis. (2005) Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank. Montserrat Civit, Sandra Kübler, and Ma. Antònia Martí, editors, Proceedings of The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005), 149-160, Barcelona, Spain, December 2005. Universitat de Barcelona.

Simov, K. and P. Osenova. (2003) Practical annotation scheme for an HPSG treebank of Bulgarian. Proc. of LINC-2003, 17–24.

van der Beek, L., G. Bouma, R. Malouf, and G. van Noord. (2002) The Alpino dependency treebank. Computational Linguistics in the Netherlands (CLIN).

Xue, Nianwen, Fei Xia, Fu-Dong Chiou, and Martha Palmer. (2005) The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2): 207-238.

References

Buch-Kromann, M. (2006) *Discontinuous Grammar. A dependency-based model of human parsing and language acquisition*. Dr.ling.merc. dissertation, Copenhagen Business School, July 2006.

Covington, M. (2001) A fundamental algorithm for dependency parsing. *Proceedings of the 39th Annual ACM Southeast Conference*, Athens, Georgia, 2001.

Covington, M. (2003) *A Free-Word-Order Dependency Parser in Prolog*. Artificial Intelligence Center, The University of Georgia.

Cowan, N. (2001) The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24, 87-185.

Cowan, N. (2005) *Working memory capacity*. Hove, East Sussex, UK: Psychology Press.

Cuetos, F. and D.C. Mitchell. (1988) Crosslinguistic differences in Parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition*, 30, 73-105.

Ferrer i Cancho, R. (2004) Euclidean distance between syntactically linked words. *Physical Review E*, 70, 056135.

Ferrer i Cancho, R. (2007) Why do syntactic links not cross? *Europhysics Letters*,

76, 1228.

Ferrer i Cancho, R. (2008) Some Limits Of Standard Linguistic Typology. *Advances in Complex Systems*, 11(3): 421–432.

Frazier, Lyn (1978) *On comprehending sentences: Syntactic parsing strategies*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Frazier, Lyn (1985) Syntactic complexity, in David R. Dowty, Lauri Karttunen, and Arnold Zwicky (eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives*, 129-189, Cambridge: Cambridge University Press.

Gibson, E. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition*, 68, 1-76.

Gibson, E. and N. J. Pearlmutter, (1998) Constraints on sentence comprehension. *Trends in Cognitive Sciences* 2, 262-268.

Gibson, E. (2000). The dependency locality theory: a distance-based theory of linguistic complexity. In: *Image, language, brain*. Marantz, A. et. al. (eds). Cambridge, MA: The MIT Press, pp. 95-126.

Grodner, D. and E. Gibson, (2005) Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 2005, 261-290.

Hawkins, J. A. (1994) *A performance theory of order and constituency*. Cambridge, England: Cambridge University Press.

Hawkins, J. A. (2004) *Efficiency and Complexity in Grammars*. Oxford University Press.

Hays, D. G. (1964) Dependency Theory: A Formalism and Some Observations. *Language* 40: 511-525.

Hellwig, P. (2006) Parsing with Dependency Grammars. in *Dependency and Valency. An International Handbook of Contemporary Research*. Edited by V. Agel, L.M. Eichinger, H.-W. Eroms, P. Hellwig, H.-J. Heringer, H. Lobin. Volume II. Mouton.

Heringer, H. J., Strecker, B., & Wimmer, R. (1980) *Syntax: Fragen-Lösungen-Alternativen*. München: Wilhelm Fink Verlag.

Hiranuma, S. (1999) Syntactic Difficulty in English and Japanese: a Textual Study. *UCL Working Papers in Linguistics*, 11, 309-322.

Hudson, R. (1990) *English Word Grammar*. Oxford: Blackwell.

Hudson, R. (1995) Measuring Syntactic Difficulty. Unpublished paper. <http://www.phon.ucl.ac.uk/home/dick/difficulty.htm> (2008-7-6)

Hudson, R. (1996) The difficulty of (so-called) self-embedded structures. *UCL Working Papers in Linguistics*. <http://www.phon.ucl.ac.uk/home/PUB/WPL/uclwpl8.html> (2008-7-6)

Hudson, R. (2004) Are determiners heads? *Functions of Language*. 11(1): 7 - 43.

Hudson, R. (2007) *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Jay, Timothy B. (2004) *The Psychology of Language*. Beijing: Beijing University Press.

Jurafsky, D. (2003) Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, (Eds.), *Probabilistic Linguistics*. Cambridge, Massachusetts: The MIT Press. p. 39-95.

Ke, J.-Y. and Y. Yao (2008) Analyzing language development from a network approach. *Journal of Quantitative Linguistics* 15 (1):70–99.

- King, J. and M. A. Just (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30, 580–602.
- Lecerf, Y. (1960) Programme des conflits - modèle des conflits. *Rapport CETIS*. No. 4, Euratom. p. 1-24.
- Liu, H. (2007) Probability distribution of dependency distance. *Glottometrics* 15: 1-12.
- Liu, H. (2008a) The complexity of Chinese dependency syntactic networks. *Physica A*. 387: 3048-3058.
- Liu, H. (2008b) A quantitative study of Chinese structures based on dependency treebanks. *Yangtze River Academic*. 3. in Chinese.
- Liu, H. and F. Hu (2008) What role does syntax play in a language network? *Europhysics Letters*. 83 (2008) 18002.
- Liu, H., R. Hudson and Zh. Feng (forthcoming). Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory*.
- Miller, George, (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Miller, G. A. and N. Chomsky (1963) Finitary Models of Language Users, in R. Luce, R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*, Ch 13, New York: Wiley.
- Nivre, J. (2006) *Inductive Dependency Parsing*. Dordrecht: Springer..
- Roelcke, T. (2002) Efficiency of communication. *Glottometrics*, 4: 27-38.
- Schubert, K. (1987) *Metataxis: contrastive dependency syntax for machine translation*. Dordrecht: Foris.
- Temperley, D. (2007) Minimization of Dependency Length in Written English. *Cognition*, 105: 300-333.
- Temperley, D. (2008) Dependency Length Minimization in Natural and Artificial Languages. *Journal of Quantitative Linguistics*, 15: 256-282.
- Tesnière, L. (1959) *Eléments de la syntaxe structurale*. Paris: Klincksieck.
- Weckerly, J., & Elman, J.L. (1992). A PDP approach to processing center-embedded sentences. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
- Yngve, V. (1960) A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, 104: 444-466.
- Yngve, V. (1996) *From Grammar to Science: New Foundations for General Linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing Company.