The properties of rare and complex syntactic constructions in English A corpus-based comparative study

Ruochen Niu
Zhejiang University
China
niuruochen@126.com

Yaqin Wang
Guangdong University of Foreign Studies
China
wyq322@126.com

Haitao Liu
Zhejiang University
China
htliu@163.com

Abstract

The study adopts a corpus-based approach to investigate rare and complex constructions in English, such as it-clefts and topicalization. Two dependency-based metrics, namely dependency distance (DD) and hierarchical distance (HD) were used to measure and compare the syntactic complexities at the linear and hierarchical levels of three treebanks, i.e., one specially-designed corpus containing many difficult and infrequent constructions and two reference corpora containing normal constructions. It was found that compared to normal constructions, syntactically infrequent and complex constructions may enjoy higher complexity at the linear level, i.e., longer dependencies, but they are not necessarily more complex at the hierarchical level. In fact, the results suggest that syntactically marked constructions are less tolerant of structural or hierarchical complexity, which may be motivated by a mechanism to avoid self-embedding or recursion driven by limited cognitive resources of human beings.

1 Introduction

Complex and rare syntactic constructions, such as it-clefs, subject-extracted relative clauses and topicalization, constitute an important part of English languages. Generally, such a syntactically marked construction has one or a combination of the following characteristics: (1) a special word order that is different from the canonical word order of the language; (2) non-local dependencies; (3) crossing dependencies that violate the principle of projectivity (also referred to as "discontinuities"). Due to these unique features, these complex constructions have been a central topic of study in many fields of linguistics: in psycholinguistics, they are known for being difficult to process, corresponding to longer reading times in language processing experiments (e.g., Grodner and Gibson, 2005); in syntax, they pose great challenges to grammarians attempting to describe and theorize their structures (e.g., Hudson, 2010; Osborne, 2019). While the above studies are conducive to our understanding of the complex and rare constructions, they are limited in two aspects: (1) the materials used to draw conclusions are often limited in number and range; (2) different constructions are studied independently as individual phenomena. This has hindered our understanding of complex constructions as a whole.

To address the above issue, the current study adopts a corpus-based approach to analyzing the properties of syntactically complex and rare constructions in English. Three corpora were adopted, one specially designed to contain as many of these hard-to-process constructions as possible (Futrell et al., 2021) and two self-built reference corpora sampled from the British National Corpus (Burnard, 2000). Comparisons were drawn on the syntactic complexities of the treebanks at the linear and hierarchical dimensions, which were measured by *dependency distance* (DD) (e.g., Hudson, 1995; Ferrer-i-Cancho, 2004; Liu, 2008; Liu et al., 2017) and *hierarchical distance* (HD) (e.g., Jing and Liu, 2015; Liu and Jing, 2016; Komori et al., 2019), respectively. By comparing these metrics and their distributions in the three treebanks, we are able to gain insights into the structural properties of complex and rare constructions in English and natural languages at large. The following of the manuscript is organized as follows: Section 2 and Section 3 introduces the methods and materials, Section 4 reports and discusses the results, and Section 5 draws a conclusion.

2 Dependency Distance and Hierarchical Distance

This section defines and illustrates the syntactic complexity measures we used to make comparisons among the corpora. They are dependency distance (DD) and hierarchical distance (HD) from the theoretical framework of dependency grammar.

2.1 Two dimensions of a syntactic tree

The study adopts *dependency grammar* as opposed to constituency grammar to analyzing syntactic structure. Dependency grammar views sentence structure as composed of direct links between words, i.e., dependencies (e.g., Tesnière, 1959/2015; Heringer et al., 1980; Mel'čuk, 1988; Hudson, 2010; Nivre, 2015; Osborne, 2019). Between the two words building a dependency relation, the word that expresses the core meaning or licenses the appearance of the other word is called the *head* (or governor), and the word that complements or modifies is the *dependent* (or subordinator).

A dependency structure of a sentence can be shown by a two-dimensional tree that clearly illustrates the two ordering principles. Between them, the horizontal or *x* axis represents the linear order of words that is based on the left-to-right occurrence and the vertical or *y* axis the hierarchical order of words according to the head-dependent relation (c.f., Tesnière, 1959/2015; Osborne, 2019). To illustrate, the dependency tree of an example sentence is given as Figure 1:

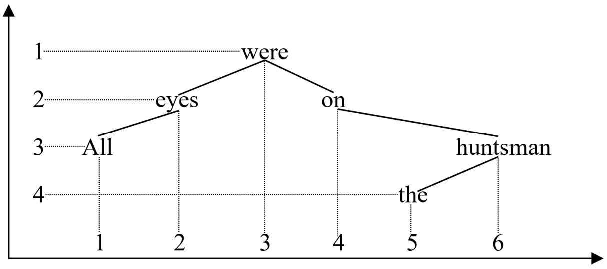

Figure 1. A Two-Dimensional Dependency Tree

In Figure 1, words are connected by concrete lines representing dependencies. Each word projects upon two axes, and the numbers of a word at the *x* and *y* axes stand for its linear and hierarchical orders within the sentence, respectively. Note that we define the hierarchical order of the sentence root, i.e., "were", as 1. As a result, the hierarchy of its dependents, i.e., "eyes" and "one", is 2.

2.2 DD and HD

Based on the above background, DD and HD are two metrics proposed and used for the measurement of the syntactic complexity at the linear and hierarchical level, respectively (e.g., Hudson, 1995; Ferrer-i-Cancho, 2004; Liu, 2008; Jing and Liu, 2015; Komori et al., 2019). Their definitions are given as follows:

DD
The absolute value of the linear order difference between a head and its dependent. [1]
HD
The hierarchical order difference between a word and the sentence root.
MDD
The mean of DD of a sample, e.g., a treebank.
MHD
The mean of HD of a sample, e.g., a sentence.

To give an example, there are six words and five dependencies in the sentence shown in Figure 1.

[1] In previous studies (e.g., Jiang and Liu, 2015; Wang and Liu, 2017), DD is a directed measure by which a positive and negative value denotes a head-final and head-initial relation, respectively; but it is always the absolute value of DD that is used when measuring the syntactic complexity. As the current study is not concerned with head-dependent ordering, we simply define DD as its absolute value to avoid confusion.

The DDs for these dependencies are: 1 (between “All” and “eyes”), 1 (between “eyes” and “were”), 1 between “were” and “on”), 2 (between “on” and “huntsman”) and 1 (between “the” and “huntsman”). Thus, the MDD of the sentence is their mean, i.e., 1.2. In the meantime, the HDs for “All”, “eyes”, “on”, “the” and “huntsman” are 2, 1, 1, 3, and 2, respectively. Thus, the MHD of the sentence is their mean, i.e., 1.8.

The motivation of using DD and HD as syntactic complexity metrics is related to the general cognitive constraints underlying language processing. While DD has been found to be related to working memory capacity limits (see Liu et al., 2017 for a review), HD has been associated with the decay of spreading activation (e.g., Hudson, 2010; Jing and Liu, 2015). Similar metrics have also been proposed and acknowledged within other syntactic frameworks (e.g., Yngve, 1960; Köhler and Altmann, 2000; Hawkins, 2004; Grodner and Gibson, 2005).

3 Dependency Treebanks

This section introduces the dependency-annotated corpora, i.e., treebanks, used in our study. They are the Natural Stories Corpus (hereafter abbreviated as the NS Corpus) that contain many rare and complex syntactic constructions (Futrell et al., 2021) and the two reference treebanks that we built based on the British National Corpus (Burnard, 2000).

3.1 Natural Stories Corpus (NS)

The NS Corpus is a “constructed-natural” corpus of English (Shain et al., 2016), i.e., it is designed to contain many infrequent and hard-to-process syntactic constructions while still sounding fluent to native speakers. Including a high proportion of syntactically marked constructions—non-local VP conjunction, it-cleft, topicalization, etc., the corpus is suitable for exploring the features of complex and rare syntactic constructions.

The NS Corpus is composed of ten edited children’s stories, e.g., the Bradford’s Boar. It provides three types of hand-corrected syntactic parses by Stanford Parser, from which we adopted the Stanford Dependencies parses and the Penn style PoS tags. Before data analysis, a thorough consistency check of all the parses was conducted. The trimmed treebank contains 464 sentences and 10,257 word tokens in total. Figure 2 illustrates the format of the treebank:

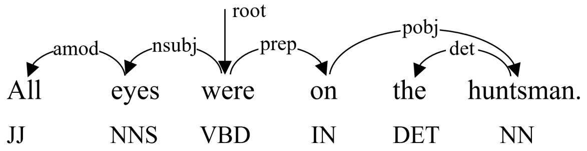

Figure 2. Format of the Treebank

In which the dependency relation is represented by the directed arc pointing from the head to the dependent. Upon each arc marks the type of the dependency, with *amod* denoting adjective modifier, *nsubj* noun subject, *prep* preposition, *pobj* object of a preposition, *det* determiner and *root* sentence root. Below each word is the word’s category, with *JJ* standing for adjective, *NNS* plural noun, *VBD* verb in past tense, *IN* preposition, *DET* determiner and *NN* singular noun. This information provides a basis for further analysis.

3.2 British National Corpus (IMA and INFO)

The comparable corpora were built on the British National Corpus (also called the BNC Corpus), which is a corpus of contemporary English containing a variety of domains. BNC’s written component (as opposed to spoken) can be largely divided into two genres, namely imaginative (hereafter abbreviated as the IMA Corpus), which is composed of fiction or other literary works, and informative (hereafter abbreviated as the INFO Corpus), which includes a variety of domains, e.g., science, word affairs and leisure (Burnard, 2000).

To build the reference treebanks, i.e., IMA and INFO Corpus, we randomly selected approximately 100,000 word tokens for each of the two genres from the BNC Corpus. Then, these two corpora were automatically annotated using Stanford Parser (version 3.9.2, de Marneffe et al., 2006) and hand-corrected by experienced annotators. The standardized IMA and INFO treebanks enjoy the same annotation scheme as the NS Corpus shown in Figure 2, but they are much larger in size—the IMA Corpus has 103,171 word tokens in 7,669 sentences and the INFO Corpus has 120,974 word tokens in 6,471 sentences. We think it is appropriate to have two comparable corpora given the potential effects of genre on syntactic complexity measures (e.g., Wang and Liu, 2017). Between the two treebanks we built, the IMA Corpus has the similar genre to the NS Corpus, and the INFO Corpus is a more general corpus of a wider coverage of the language which makes it a more suitable reference corpus (Leech, 2002).

3.3 Quantitative Properties of the Treebanks

A preliminary analysis was conducted to obtain a quick overview of the treebanks. Properties such as *mean sentence length* (MSL), *mean dependency distance* (MDD) and *mean hierarchical distance* (MHD) were focused. The results are presented in Table 1:

	Word Tokens	Sentences	MSL	MDD	MHD
NS	10257	464	22.1034	2.5719	3.0789
IMA	103171	7669	13.4530	2.2614	3.1030
INFO	120974	6471	18.6948	2.3466	3.2841

Table 1. Quantitative Properties of the Treebanks

In which *sentence length* (SL) is measured by the number of words in a sentence, and MSL the mean of all SL of a corpus. From Table 1, it is clear that (1) the MDD and MHD of the three treebanks are all below 4. This is supportive of previous findings proposing a threshold of MDD and MHD equal to working memory capacity limits (e.g., Liu, 2008; Liu and Jing, 2016); (2) the NS Corpus has a greater MSL than the INFO and IMA Corpus. This means that the three treebanks are not suitable for direct comparison because longer sentences usually lead to larger MDD and MHD (Jiang and Liu, 2015; Jing and Liu, 2015); (3) despite a larger MSL, NS has a smaller MHD than the other two treebanks, which is contradictory to our expectation and deserves more investigation.

4 Results and Discussion

In Section 3.3, we found that the mean sentence length (MSL) of the NS Corpus is greater than that of the IMA and INFO Corpus. To control for the effects of SL on the results, four SL groups were selected based on the distribution of SLs in the treebanks.[2] This section reports and discusses the findings in comparing the DD and HD related properties of the three treebanks at different sentence lengths.

4.1 DD Distribution

In previous studies, the DD distributions of natural languages have been revealed to exhibit a long tail, which indicates a preference for short dependencies driven by limited capacity of working memory (e.g., Liu, 2008; Jiang and Liu, 2015; Wang and Liu, 2017). To investigate whether the NS Corpus demonstrates any universalities or peculiarities in this regard, the frequencies of the dependencies at different

[2] Because the NS Corpus is small in size, we have to make sure that after controlling for SL it still has enough data for analysis. The selected SL groups are therefore the top four groups of SL that have the most number of sentences in the NS Corpus, i.e., (sentences made of)16-20 (words), 21-25, 11-16 and 26-30.

DD for different SL groups were extracted from the three treebanks; these frequencies were then transformed to corresponding probabilities (or proportions) for comparison among the treebanks.[3] The results are shown in Figure 3:

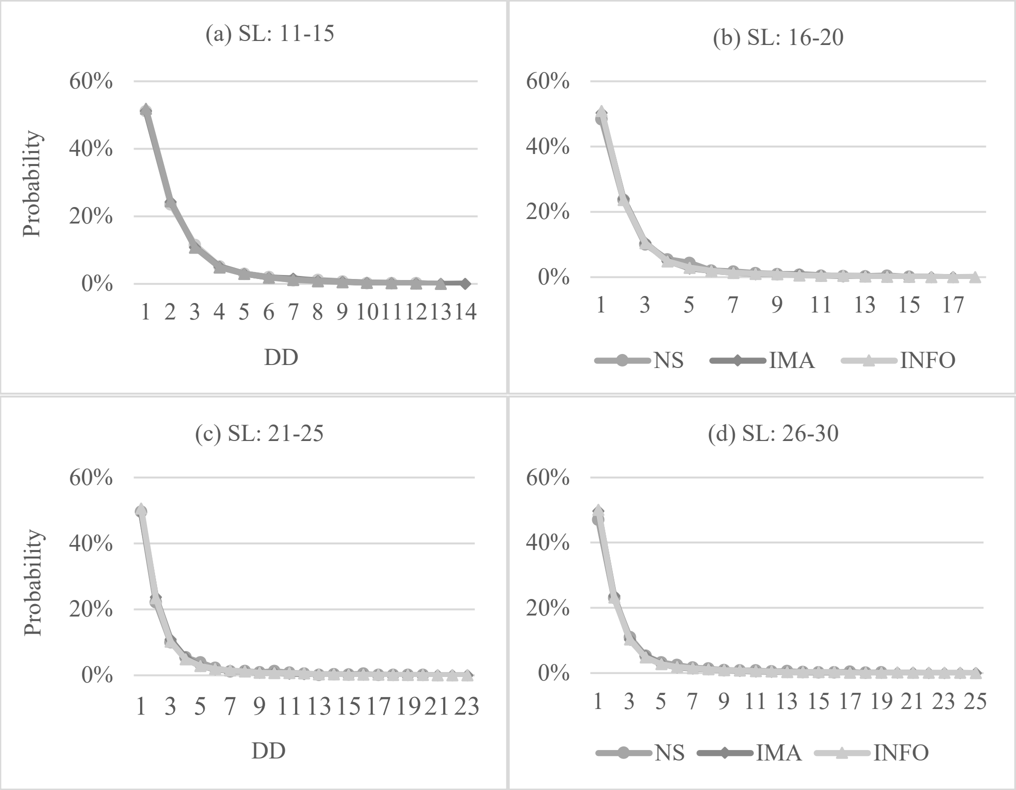

Figure 3. DD Distribution of the Treebanks at Varying SLs

Figure 3 shows that for all four sentence length (SL) groups, the dependency distance (DD) distributions of the three corpora exhibit a similar, almost identical long tail, i.e., the proportion of dependencies is the highest when DD = 1, and drops significantly when DD is increased. This indicates that (1) the preference for short dependencies is not affected by sentence lengths, which corroborates previous findings (Jiang and Liu, 2015); (2) as a corpus that includes many complex syntactic constructions, the NS Corpus is not distinctly different from the reference corpora containing normal constructions in terms of the DD distributions. These findings provide further support to the account that the preference to minimize DD is a language universal driven by general cognitive constraints of human beings rather than intra-linguistic factors (c.f., Liu et al., 2017).

4.2 HD Distribution

In Section 4.1, we found that the NS Corpus is not different from the two reference corpora in terms of the tendency to minimize syntactic difficulty at the linear level, i.e., DD. In this section, we explore their potential similarities and differences at the hierarchical level, i.e., in terms of the HD distributions. Unlike the DD distribution, the HD distribution of natural languages has attracted little attention from the academia. The only exception is Liu (2017) who studied the distribution of hierarchies in three languages and attributed the universalities found to the valency patterns in natural languages proposed by Tesnière (1959/2015). Since HD has been used as a syntactic complexity metric at the hierarchical level in both

[3] Because the treebanks vary greatly in terms of size, it is preferable to use probabilities (or proportions) rather than frequencies for comparison. The probability of the dependencies at a given DD is calculated by dividing the real frequency of dependencies at that DD in a treebank by the total frequency of all dependencies under such circumstance. For example, when SL is between 11 and 15, the frequency of the dependencies with a DD of 1 is 480 in the NS Corpus, and the total frequency of the dependencies with all possible DDs is 938. Thus, the probability of the dependencies at DD = 1 in this case is 480/938 = 51.17%, as shown in Figure 3(a).

general and applied linguistics (e.g., Komori et al., 2019), direct investigation into its distribution may yield new insights into the hierarchical complexities of human languages.

To obtain the HD distributions of the three treebanks, frequency data at different HD for the four sentence length (SL) groups were extracted from the three corpora and transformed into probabilities for cross-corpus comparison. The results are illustrated using Figure 4:

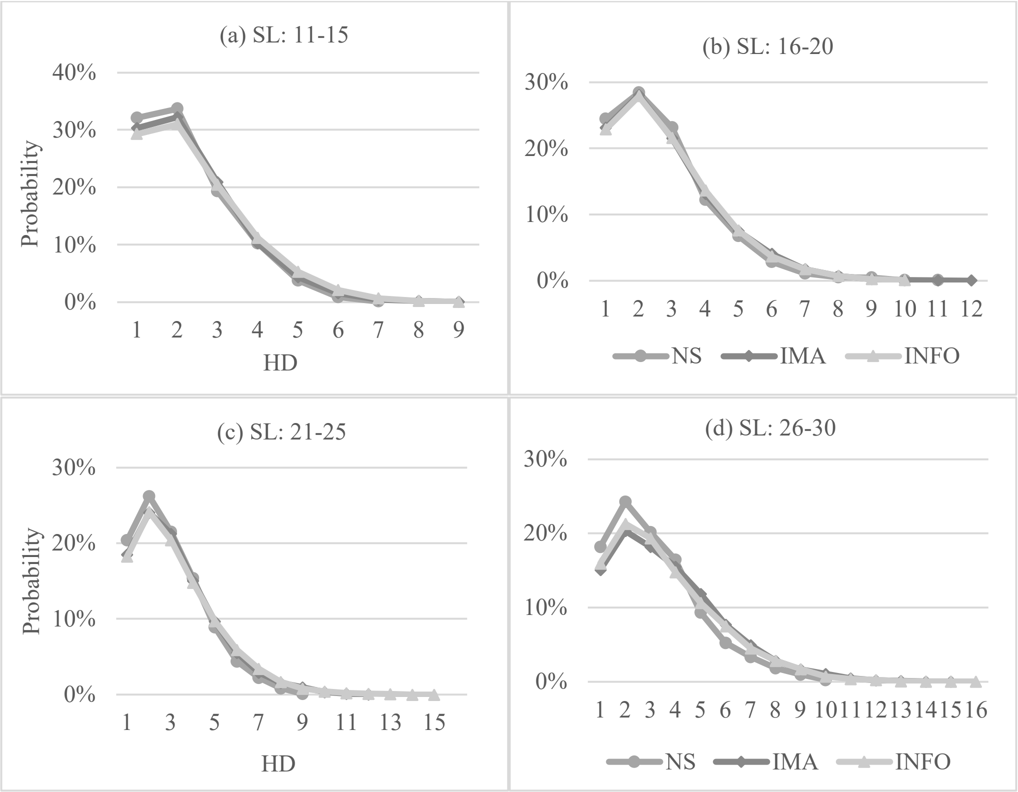

Figure 4. HD Distribution of the Treebanks of Varying SLs

It was found that (1) for all SLs and treebanks, the probability (of words at a given HD) increases when HD is increased from 1 to 2, reaches the peak when HD = 2, and then decreases sharply when HD keeps increasing. These results accord well with Liu (2017)'s findings using another metric called hierarchy (which equals to HD + 1), and suggests that the hierarchical syntactic complexity of human languages is also constrained; (2) for all three treebanks, the proportions of shorter HD (HD ≤ 3) exhibit a decreasing trend when sentences become longer. From Figure (4a) to (4d), the peaks of the curves at HD = 2 decreases from around 30% when SL is between 11 and 15 to around 20% when SL is between 26 and 30. This indicates that unlike the DD distribution that is hardly affected by SL, the HD distribution is influenced by SL to a greater extent; (3) when SL is increased, the NS Corpus seems to exhibit a stronger tendency to avoid longer HDs compared to the two reference corpora. As shown in Figure (4a) and (4b), the HD distributions of the three treebanks are similar when SL is rather small, i.e., between 11-15 and 16-20. However, the slope of the curve for the NS Corpus seems to be steeper than the other corpora when sentences become longer; this is particularly evident in Figure (4d), in which NS has higher proportions of words with shorter HDs (HD ≤ 3) but lower proportions of words with longer HDs (HD ≥ 4).

In general, the results in this section suggest that although the three treebanks share some commonalities in terms of their HD distributions, the NS Corpus that enjoys a higher proportion of complex and infrequent syntactic constructions such as relative clause and it-clefts, seems to demonstrate a greater tendency for short HDs when SL is increased. In addition, our results indicate that unlike the DD distribution, the HD distribution is more likely to be affected by SL. To further validate the phenomenon observed above, we investigated the MDD and MHD of the three treebanks at different sentence lengths, which is reported and discussed in Section 4.3.

4.3 Relation between MDD and MHD

In 4.2, we found that while the three treebanks share some similarities in their HD distributions, the NS Corpus seems to have a stronger tendency to avoid longer HDs when SL is increased. This indicates that complex and rare syntactic constructions may have a stronger tendency to avoid complexity at the hierarchical level compared to normal syntactic constructions. This subsection further explores the phenomenon by examining the relation between MDD and MHD of the three treebanks.

To begin with, we calculated the MSL, MHD and MDD of the three treebanks for the above-mentioned sentence lengths groups.[4] The results are presented in Table 2:

	SL group	NS	IMA	INFO
MSL	11-15	13.2078	12.8437	13.0315
MHD	11-15	2.2281	2.3317	2.4201
MDD	11-15	2.1098	2.0903	2.0509
MSL	16-20	18.0968	17.795	17.8541
MHD	16-20	2.6728	2.7805	2.7859
MDD	16-20	2.3808	2.2601	2.2491
MSL	21-25	22.7386	22.8515	22.9785
MHD	21-25	2.9314	3.1844	3.2437
MDD	21-25	2.4819	2.4023	2.3691
MSL	26-30	27.8553	27.8283	27.8601
MHD	26-30	3.2032	3.7066	3.5953
MDD	26-30	2.646	2.4582	2.4842

Table 2. MHD and MDD of the Treebanks at Different SLs

In Table 2, the MHD and MDD of all three treebanks both increase with MSL, and the increase of MSL brings more gain on MHD than on MDD. This is consistent with previous findings of English (Liu and Jing, 2016). In addition, the NS Corpus has the highest MDD and lowest MHD among the three treebanks within each SL group. This corroborates the reciprocal relationship between MDD and MHD found in previous studies (e.g., Jing and Liu, 2015), and suggests that syntactically complex structures are not necessarily more complex in both linear and hierarchical dimensions. More importantly, the difference in MHD among the NS Corpus and the two reference corpora becomes larger when SL is increased. In other words, the other two corpora seem to be less constrained in MHD than the NS Corpus when sentence length is increased. This coincides with our finding in Section 4.2, and suggests that syntactically marked structures may be more complex at the linear level, but they tend to avoid complexity at the hierarchical level compared to normal constructions.

To visualize the trade-off relation between MDD and MHD and the differences of the treebanks in this regard, we created a bubble chart based on the MDD and MHD of all sentences in the three treebanks. [5] In Figure 5, the direction of the *x* axis denotes larger MHD and the direction of the *y* axis lager MDD. It is clear that the NS Corpus has greater overall MDD (the upper part of the chart is occupied mostly by dark grey bubbles), whereas the IMA and INFO Corpus tend to have greater overall MHD (the right side of the chart is taken up mostly by grey and light gray bubbles). This is supportive of our findings above and indicates that syntactically rare and complex constructions in English may have slightly

[4] Sentence length is controlled because it has an impact on MDD and MHD (e.g., Jing and Liu, 2015; Liu and Jing, 2016). See footnote 2 for how these four SL groups were selected.

[5] A bubble chart is a variation of a scatter chart in which the data points are replaced with bubbles. In addition to the two axes, a third piece of information about the data is shown by the size of the bubbles (here the number of sentences at a given MHD or MDD). What's more, bubble charts are helpful for analyzing the trend of the data.

longer dependencies, but they are not more complex hierarchically. In fact, our results suggest the reverse situation, i.e., syntactically complex structures, e.g., it-clefts and subject-extracted relative clauses, are less lenient on the complexity at the hierarchical level compared to normal structures.

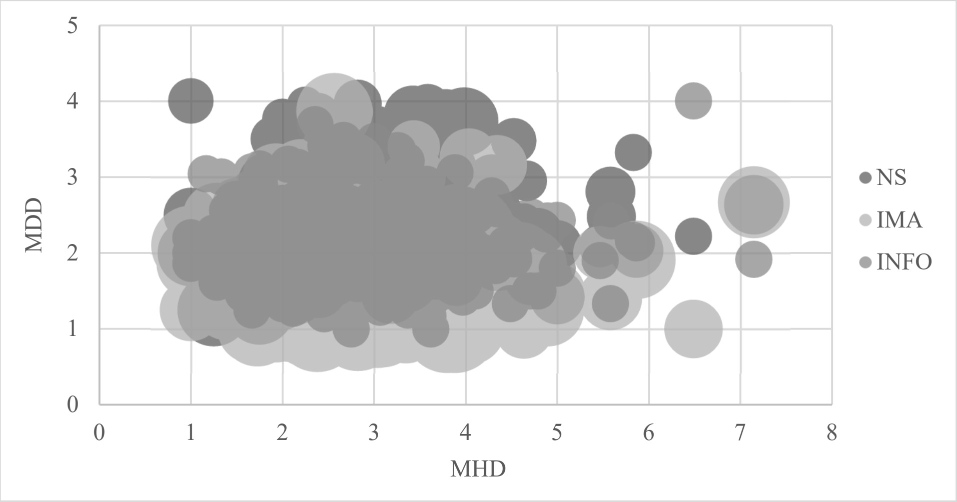

Figure 5. Reciprocal Relation between MDD and MHD in the Treebanks

Previous studies have related a word's difficulty at the hierarchical level (or HD) to the decay of spreading activation from the sentence root to that word (Jing and Liu, 2015; c.f., Hudson, 2010). This, however, does not explain why syntactically marked constructions as a whole are less lenient on the hierarchical complexity than normal constructions. After examining the distributions of grammatical relations at different HDs of the three treebanks, we found that this property of the NS Corpus may be motivated by a mechanism to avoid self-embedding, i.e., embedding of structures of the similar kind. Self-embedding (or recursion) is known as a fundamental property of human languages (e.g., Hauser et al., 2002), but researchers from different fields have found constraints on its use in natural languages, driven perhaps by limited cognitive resources of human beings (e.g., Karlsson, 2007; Christiansen and MacDonald, 2009). In other words, self-embedded structures are cognitively challenging. In our study, the syntactically complex and rare constructions in the NS Corpus have been found to be more difficult to process at the linear level. Presumably, the embedding of these constructions will induce more cognitive effort than that of normal constructions, which can at times lead to processing breakdown. Thus, the stricter constraint on the hierarchical complexity of the NS Corpus may be an attempt to avoid self-embedding to counteract the cognitive burden imposed by the presence of rare and complex constructions. In general, the above findings and discussions support the idea of language as a human-driven complex adaptive system (e.g., Christiansen and MacDonald, 2009; Liu, 2018).

5 Conclusion

Our study investigates the properties of syntactically marked constructions based on a specially designed corpus containing many infrequent and complex constructions and two reference corpora in the same annotation scheme. By examining and comparing their syntactic complexities at the linear and hierarchical levels (measured by DD and HD, respectively), we found that syntactically complex constructions may be more difficult to process at the linear level, but their hierarchical structures are not necessarily more complex than normal constructions. We attribute the stricter constraint on the hierarchical complexity of the complex constructions to an attempt to avoid self-embedding or recursion in natural language processing, which is ultimately motivated by limited cognitive resources. Altogether these findings indicate properties of natural languages as a human-driven self-adaptive complex system, which calls for more interdisciplinary research.

Acknowledgements

We are grateful to the three anonymous reviewers, whose insightful comments helped improve the content of the manuscript.

References

Lou Burnard (ed.). 2000. *Users Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Morten H. Christiansen and Maryellen C. MacDonald. 2009. A usage-based approach to recursion in sentence processing. *Language Learning*, 59: 126–161.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454.

Ramon Ferrer-i-Cancho. 2004. Euclidean distance between syntactically linked words. *Physical Review E*, 70: 056135.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55: 63–77.

Daniel. J. Grodner and Edward Gibson. 2005. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29: 261–290.

Marc D. Hauser, Noam Chomsky, and W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298: 1569–1579.

John A. Hawkins. 2004. *Efficiency and Complexity in Grammar*. Oxford: Oxford University Press.

Hans J. Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. München: Wilhelm Fink Verlag.

Richard Hudson. 1995. Measuring syntactic difficulty. Unpublished paper. Available at http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf (Accessed 5 June 2019).

Richard Hudson. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.

Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English-Chinese dependency treebank. *Language Sciences*, 50: 93–104.

Yingqi Jing and Haitao Liu. 2015. Mean hierarchical distance: Augmenting mean dependency distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 161–170.

Fred Karlsson. 2007. Constraints on multiple center-embedding of clauses. *Journal of Linguistics*, 43: 365–392.

Reinhard Köhler and Gabriel Altmann. 2000. Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics*, 7(3): 189–200.

Saeko Komori, Masatoshi Sugiura, and Wenping Li. 2019. Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data. In *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019)*, pages 130–135.

Geoffrey Leech. 2002. The importance of reference corpora. In *Hizkuntza-corpusak. Oraina eta geroa*. Available at https://uzei.eus/online/dokumentazioa/biltzarrak/corpus-jardunaldia-2002/(Accessed 26 September).

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9: 159–191.

Haitao Liu. 2017. *Juzi jiegou cengji de fenbu guilv* "The hierarchical distribution of sentence structure". *Foreign Language Teaching and Research*, 49(3): 345–352.

Haitao Liu. 2018. Language as a human-driven complex adaptive system. *Physics of Life Reviews*, 26-27: 149–151.

Haitao Liu and Yingqi Jing. 2016. *Yingyu juzi cengji jiegou jiliang fenxi* "A quantitative analysis of English hierarchical structure". *Journal of Foreign Languages*, 6: 2–11.

Haitao Liu, Chunshan Xu, and Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21: 171–193.

Igor A. Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press, Albany.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2015, Part I, LNCS 9041*), pages 3–16.

Timothy Osborne. 2019. *A Dependency Grammar of English. An Introduction and Beyond*. Amsterdam: John Benjamins.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC 2016)*, pages 49–58.

Lucien Tesnière. 2015. *Elements of Structural Syntax* (original work published in 1959, translated by Timothy Osborne and Sylvain Kahane). Amsterdam: John Benjamins.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59(866): 135–147.

Victor H. Yngve. 1960. A model and an hypothesis for language structure. In *Proceedings of the American Philosophical Society (Vol. 104, No. 5)*, pages 444–466.