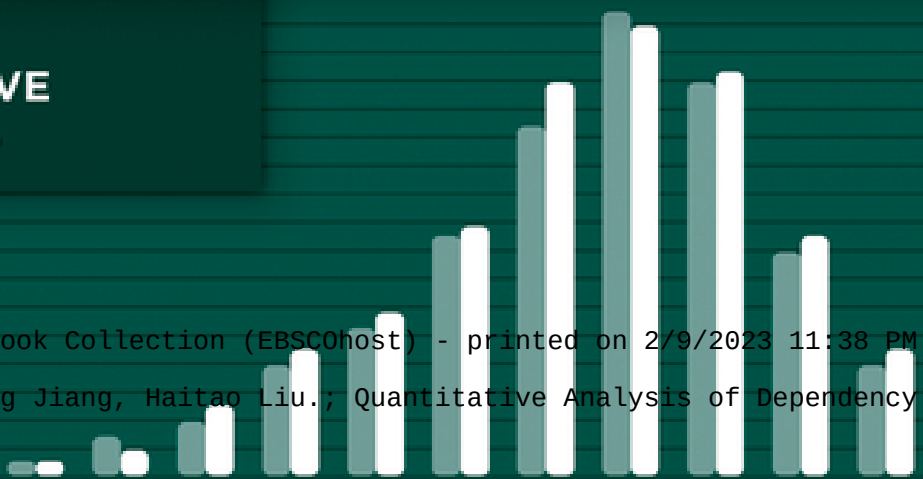
DE GRUYTER
MOUTON

Jingyang Jiang, Haitao Liu (Eds.)

QUANTITATIVE ANALYSIS OF DEPENDENCY STRUCTURES

QUANTITATIVE LINGUISTICS



Jingyang Jiang and Haitao Liu (Eds.)
Quantitative Analysis of Dependency Structures

Quantitative Linguistics

Editor
Reinhard Köhler

Volume 72

Quantitative Analysis of Dependency Structures

Edited by
Jingyang Jiang
Haitao Liu

DE GRUYTER
MOUTON

ISBN 978-3-11-056577-5
e-ISBN (PDF) 978-3-11-057356-5
e-ISBN (EPUB) 978-3-11-057109-7
ISSN 0179-3616

Library of Congress Control Number: 2018956317

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing and binding: CPI books GmbH, Leck

www.degruyter.com

Preface

The practice of using syntactic dependency to analyze the structures of human languages enjoys a long history. Ranging from the Paninian grammar of ancient India, the Modistae syntactic theory in the Middle Ages of Europe, the medieval Arabic grammar to the traditional grammars of many countries in the world, there infiltrates more or less the idea of syntactic dependency.

Since the English Treebank of the University of Pennsylvania was launched in 1993, (computational) linguists around the world have set off an upsurge of "building treebanks". Twenty years of research and practice have shown that not only the "tree" types in the treebanks tend to change from phrase structure to dependency structure, but also the latter has the potential to virtually dominate the annotation schemes of all treebanks.

The thriving of dependency analysis in the (computational) linguistic field may be attributed to the following advantages: more convenient to convert from the syntactic level to the semantic level; more suitable for processing the languages with free word order; more psychological reality; more applicable to some fields in natural language understanding; and easier to construct high-precision syntactic parsers based on machine learning.

Despite that dependency analysis is enjoying tremendous popularity in the real world, we rarely witness its existence in the linguistics (syntax) classrooms and textbooks in universities. This disconnection between theory and practice is not a preferable phenomenon in any sense.

French linguist Lucien Tesnière is recognized as the founder of modern dependency grammar. But owing to its long history and various sources, dependency grammar has evolved to be a sentence analysis method which is much more open and has more variants. Consequently, the discussion of dependency grammar will inevitably involve a considerable number of existing analysis methods used to annotate authentic language data. Although these methods are of different origins and forms, the syntactic dependencies they use generally have the following three features: binary, asymmetry, and labelledness. It is these commonalities that lay the foundation for the current volume.

The discovery of linguistic laws is a necessary step in the formation of linguistic theories. Therefore, quantitative linguistic (QL) research based on the corpora annotated with dependency structure has important implications for the formation of a scientific syntactic theory. Although we already have some sporadic quantitative studies on dependency structures, the depth and breadth are far from satisfying the needs of the real world. The emergence of a large number of multi-language dependency treebanks provides valuable language

https://doi.org/10.1515/9783110573565-201

resources for us to conduct more in-depth quantitative research into dependency structure, and also enables us to carry out studies, such as investigations on language cognition and linguistic typology based on distance and direction between two interrelated words, which are not able to be conducted if only based on Tesnière's original schema. Hence, this book is a welcome product that not only inherits Tesnière's ideas but also integrates all these newly-introduced concepts in the age of Artificial Intelligence (AI). By collecting 16 related papers from 32 authors that apply quantitative methods, the volume explores many aspects of dependency relations from various perspectives.

Richard Hudson uses Google N-grams to present the historical changes of some concepts related to dependency analysis, and puts forward his own views on the following issues: the relationship between corpus and language systems as well as language processing, the psychological reality of dependency structure and phrase structure, syntactic tree representation and network structure, projectivity and word order, etc. Strictly speaking, his article is not a QL article in a general sense, but the topics he touches on are closely related to most of the research in this volume.

The basis of dependency analysis is the dependency relationship between two words. Dependency is an asymmetric relationship of syntactic functions, which makes the sentence structure derived from the dependency analysis take the form of a tree structure. We can thus use quantitative methods to study some of the formal characteristics of the tree structures, as well as the relationship between structure and function.

Hongxin Zhang and **Haitao Liu** use dependency treebanks of two different languages to examine the interrelations among dependency tree widths, heights and sentence lengths, thus taking a step further toward constructing a synergetic syntactic model based on dependency structure. **Radek Čech** and his colleagues introduce a new quantitative unit of linguistic structure—dependency frame, and adopt the Czech Universal Dependency Treebank to study the rank-frequency distribution regularity of the dependency frame and the relationship between this unit and some particular syntactic functions. **Anat Ninio** converts the treebanks of Hebrew parents' and children's speech into a bipartite network and uses this network to study the relationship between syntax and communicative functions. Her research shows that syntactic structures and communicative functions are not fused in such a way as the construction grammar contends.

Dependency is a kind of inter-word relationship, whose foundation lies in the fact that every word has a potential capacity to combine with other words, i. e. valency. Traditionally, the valence of a word can be looked up in the diction-

ary, and this is called the static valency. Once a word enters a sentence, this potential combining ability is realized and a dependency relation is thus formed. In other words, we can use the dependency treebank to study the combining power of words. Under this circumstance, the valency becomes a dynamic one.

Andrei Beliankou and **Reinhard Köhler** extract valency structures from a Russian dependency treebank, and investigate its frequency distribution. Contrary to many traditional valency studies, the research once again negates, at least from a theoretical point of view, the necessity of making a distinction between complements and adjuncts from a quantitative perspective. **Huiyuan Jin** and **Haitao Liu** use a spoken Chinese dependency treebank to explore the regularities between verb valency and ellipsis from a dynamic perspective, which enables us to have a more objective understanding of the effect of valency in authentic communication contexts. **Haruko Sanada** utilizes the Japanese valency database to study, from the perspective of dependency types and part of speech, the change of information amount of dependency type in different positions of Japanese clauses. The study provides us with a further understanding of the combining power of words in different positions of sentences. **Qian Lu**, **Yanni Lin**, and **Haitao Liu** use a Chinese dependency treebank and an English one to dig into the relationship between dynamic valency and the linear distance between two words with a dependency relationship. Their study integrates words' combining power (dynamic valency), network structure, and human language processing mechanisms together, which echoes with many other articles in this book.

As the last sentence of Hudson's article goes, "dependency analysis of corpora can throw considerable light on cognition." The reason for this statement is that, on the one hand, there have been many cognitive experiments confirming that dependency analysis has better psychological reality; on the other hand, people have found some syntactic measures that can reflect the constraints of cognitive mechanisms on language structure, the most important one from these measures being the dependency distance which refers to the linear distance between two syntactically related words in a sentence. The main reason that dependency distance can be associated with human cognitive mechanisms is that in the process of incremental parsing from left to right, if the two words with syntactic relations are so far apart that they are beyond the working memory capacity, then it may cause some difficulties in understanding. Subsequently, a large number of related studies have emerged to explore various factors that influence the dependency distance, among which the most noteworthy finding might be the minimization of dependency distance. What this

argues is that when people are making sentences, they are more inclined to choose certain kinds of linear arrangement of words in a sentence. In such a linear arrangement, the sum of linear distances (dependency distances) between words with syntactic relations has the propensity to be minimized. However, as Hudson mentions in his article, logically speaking, the data in the corpus come from the speakers. Using such corpus to study the hearer's language processing mechanism should be handled with extra caution. It is our belief, though, that any specific issue should be specifically treated. The relation between dependency distance and human cognition demonstrates the constraints of the working memory capacity, which should not only apply to the hearer but also to the speaker. This viewpoint has been confirmed by cognitive experiments and numerous syntactic structure analyses. In recent years, through large-scale multilingual authentic corpora, it has been found that minimizing dependency distance is a universal feature of human language, a manifestation of the principle of least effort in language processing at the syntactic level. Many studies in quantitative linguistics have shown that the statistical laws of human natural language systems reflect, to a great extent, the balance and stability between a speaker and a hearer. The probability distribution of dependency distance of many languages and text types displays the same balance and stability. In other words, in most cases, the utterances of a speaker are consistent with the minimization of the dependency distance. This not only means that the utterances are constrained by the working memory capacity, but also demonstrates that in order to communicate with the listener smoothly, the speaker is able to subconsciously take the cognitive needs of the listener into account. Of course, we fully acknowledge that the relationship between language structure and cognition is very complicated. Therefore, the issues related to dependency distance also require further explorations both in theory and in practice from various perspectives. This may account for the reasons why this volume has many articles addressing this issue.

Jingyang Jiang and **Jinghui Ouyang** use the corpus of Chinese learners' written English across nine grades to study the changing patterns of the dependency distance of interlanguage. Their research finds that compared with the random languages, the interlanguage in different stages has a tendency of dependency distance minimizing, and the probability distribution parameters of the dependency distance can be used to differentiate the language level of learners. **Jingqi Yan** uses a similar method to study nine grades of deaf and hard-of-hearing students learning Chinese written language. The results show that the distribution of dependency distances of learners in different learning stages conforms to certain rules, and their parameters can also reflect the learn-

ers' proficiency level to some extent. The above two studies have not only broadened the field where minimizing dependence distance can be applied, but also demonstrated once again that the dependency distance minimizing is universal to language in that it is constrained by the general mechanism of human cognition.

The calculation of dependency distance relies on the positions of two words with syntactic relations in a sentence. Therefore, it is natural that we may wonder whether there is any regularity in the change of distance between words in different positions of a sentence. **Hua Wang** investigates this issue using a Chinese dependency treebank and finds that the mean dependency distances between words at sentence-initial and sentence-final positions are the largest, while distribution of dependency distance in each sentential position can be captured by power law. **Jinlu Liu** and **Gaiying Chai** use the self-built English treebank to study the dependency traits of English relative clauses, including their relations between dependency distance, dependency direction, the embedding position, and the length of relative clauses.

Minimizing dependency distance is a statistical feature, but in the actual language use, sometimes a long dependency will appear because the speaker needs to achieve his/her other more important intentions. But in order to reduce the difficulties in comprehending or generating a sentence, the speaker may use some language devices to reduce the difficulty of processing. Departing from this point of view, **Chunshan Xu** studies the subject post-modifiers and object post-modifiers in English. He discovers that the post-modifiers of subjects tend to be shorter than those of objects, and people also prefer to use punctuation marks to reduce the dependency distance of such subjects.

The quantitative research of dependency structures can give us an in-depth theoretical insight of the structural regularities of sentences. On top of that, it should also help us to solve some problems in applied linguistics, for example, the language acquisition problems mentioned before. If these regularities are found in authentic texts (language use), then they may also be used in areas such as the automatic classification of languages, the study of genres, and the identification of authorship. **Xinying Chen** and **Kim Gerdes** use Universal Dependency treebanks to classify more than 40 languages. Their research in typology has not only expanded our previous understanding of this field with more languages, but also enabled us to recognize the distinct difference between using syntactic-annotated treebanks and semantic-annotated treebanks when it comes to solving language classification problems. **Yaqin Wang** and **Jianwei Yan** investigate the literary materials in the LOB corpus in the hope of uncovering the relationship between different genres of literary works and their

dependency structures. Their research has not only introduced syntactic quantitative methods into the field of literary studies, but has also deepened our understanding of the relationship between genres and syntactic structures. **Alexander Mehler** and his colleagues, based on the previous research on the dependency tree, creatively propose a multidimensional model with 46 features. This model can provide information about authorship at the text level, laying a good foundation for achieving the ultimate goal of predicting the author of sentences using dependency tree structure.

It is hoped that these studies will not only help readers understand the regularities of dependency syntactic structures from the quantitative perspective, but also deepen the understanding of the relationship between dependency and cognitive mechanism. Besides, readers will be more acquainted with its application in the fields of natural language processing, language typology, language acquisition, genre analysis and many other related fields.

We are grateful to all the authors and reviewers for their cooperation and help, and we would also like to thank **Yalan Wang** for his painstaking work in typesetting and editorial work. Our most sincere thanks also go to **Reinhard Köhler**, the editor-in-chief of the *Quantitative Linguistics* series, for his advice and support at various stages.

Finally, we would like to acknowledge the National Social Science Foundation of China (*Syntactic Development of Chinese EFL Learners Based on Dependency Treebank*, Grant No. 17AYY021), the Fundamental Research Funds for the Central Universities (*Research on the Syntactic Development of English Learners; Program of Big Data PLUS Language Universals and Cognition*, Zhejiang University), and the MOE Project of the Center for Linguistics and Applied Linguistics (Guangdong University of Foreign Studies), which supported us during the preparation of this volume.

Jingyang Jiang[1], Haitao Liu[2]

1 Jingyang Jiang: Department of Linguistics, Zhejiang University, Hangzhou, China

2 Haitao Liu: Department of Linguistics, Zhejiang University, Hangzhou; Center for Linguistics & Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China

Contents

Richard Hudson

Dependency, Corpora and Cognition

Abstract: Using Google N-grams as a resource, I review the history of dependency analysis in quantitative linguistics, then address a number of general issues: (1) How corpus studies relate to cognition: I present three connections, and argue that a corpus can't be used as direct evidence either for the language system or for processing difficulty. (2) The nature of syntactic relations: I contrast n-grams with both phrase structure and dependency structure, arguing that dependency structure is most compatible with what we know about cognition. (3) The simplicity of syntactic structure: I argue that, in terms of cognitive reality, syntactic structure is too complex for simple tree diagrams and is formally a network in which words may depend on several other words. (4) The universality of syntactic structure: I argue that languages vary almost without limit, so if we respect cognitive reality we can't assume a universal set of categories or dependency patterns. On the other hand, some features are shared by some languages, so some cross-language comparison is in fact possible.

Keywords: dependency analysis; corpora and cognition; syntactic structure

1 History

Quantitative linguistics is an important tool for the study of changes through time, so I start by applying this tool to itself: the history of corpus studies. This will allow me to introduce a research tool which will be relevant throughout this paper: Google N-grams (https://books.google.com/ngrams). This facility accesses millions of books written in English (as well as a number of other languages) and dating from 1800, each classified by its year of publication. For each year, the diagram shows the relative frequency of the selected n-gram, so the figure is not affected by the size of the total corpus for that year. Fig. 1 shows the relative frequencies for the phrases *word frequency*, *quantitative linguistics* and *corpus statistics*. This graph shows that word counting has been popular since the 1920s, and remains so in spite of a small drop since the peak in the 1990s. It predates quantitative linguistics by several decades, which in turn predates the corpus age which allows corpus statistics.

Richard Hudson, University College London, London, U.K., r.hudson@ucl.ac.uk

https://doi.org/10.1515/9783110573565-001

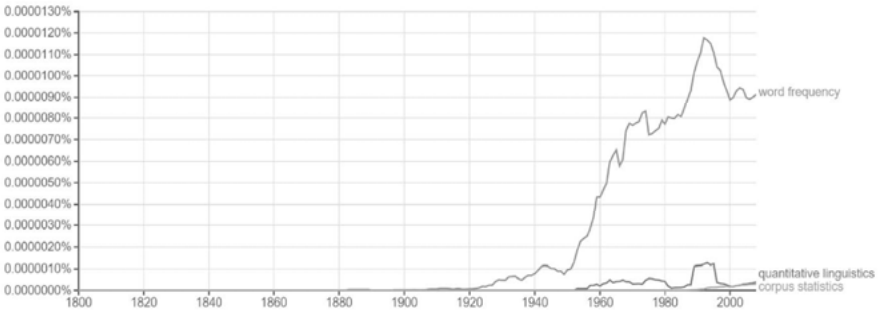


Fig. 1: The rise of word frequency, quantitative linguistics and corpus statistics

Google N-grams allows the user to zoom in on less frequent items simply by removing those that dwarf them, so the rise of corpus statistics emerges more clearly in Fig. 2 Corpus statistics is clearly a young but rising star within the firmament of quantitative linguistics.

Fig. 2: The rise of corpus statistics

This chapter is concerned with one particular area of study in quantitative, corpus-based, linguistics, namely the study of sentence structure. As we all know, the study of sentence structure predates quantitative linguistics and even word-counting, but we might not be able to give as precise a date as Google N-grams does in Fig. 3: 1882. (The Oxford English Dictionary locates the first use of *sentence structure* in 1872.) Interestingly, 1882 is also the first year for the German translation of 'sentence structure', *Satzbau*, a term which followed a similar trajectory to its English equivalent.

Fig. 3: The rise of interest in sentence structure using phrase/constituent structure and dependency structure

In contrast, the French equivalent, *structure de la phrase*, has the rather different history shown in Fig. 4. French authors were already talking about sentence structure early in the 19th century, so it's at least possible that they provided the inspiration for both German and Anglophone authors; but it was the latter that gave the idea the high profile that it's enjoyed since the early 20th century.

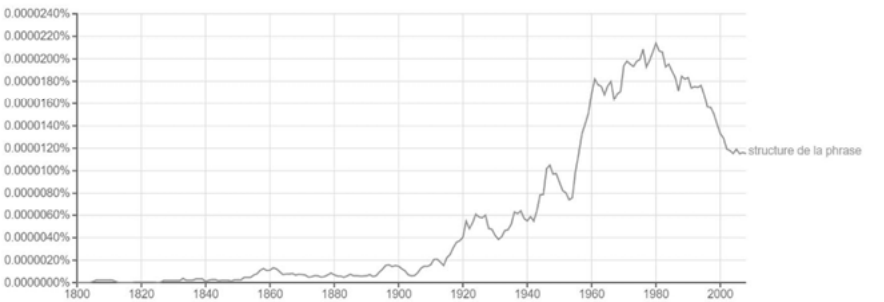


Fig. 4: The rise of sentence structure in French

We have seen that quantitative studies of texts date from the early 20th century, while the study of sentence structure is a century older. The following sections will consider how these two strands of research can, and should, be combined.

2 Corpora and cognition

Corpora relate to cognition in at least three ways:

1) as evidence for the prevailing culture.
2) as evidence for the underlying language system which is found, with some variation, in the minds of speakers.
3) as evidence for language processing.

The link to culture can be illustrated neatly by Fig. 5, which shows how the phrases *good idea* and *good man* have changed in frequency since 1800.

Fig. 5: Historical differences between 'good idea' and 'good man'

Even more interesting is the pattern in Fig. 6, comparing the frequencies of various personal pronouns. One striking feature is the difference between *he* and *she*, and another is the recent sharp rise in the first-person *I* (but not *we*). Both of these features invite discussion and explanation in terms of social and cultural changes with potentially serious implications.

However we interpret these changes, they clearly reflect profound changes in Anglophone culture, at least as represented in books. Such changes presumably have very little to do with the language system because they concern what people talk about and what they believe and think. Nevertheless, we can see that they affect the results of a corpus study.

The link between corpora and the language system is more complicated, and more directly relevant to the present book. Any language corpus is the product of human cognition, because every word in the corpus was chosen by some human mind. In that sense, then, a corpus is a window into the mind, but it doesn't tell us directly about the mind's language system because other parts

of the mind are also involved in choosing the words. For example, the human mind is prone to error, so a sentence in a corpus may be a distortion of the underlying language system: performance is an imperfect reflection of competence. Speech errors are caused by temporary lapses of memory and confusion, so by definition they won't occur in sufficient numbers to show up as n-grams (and in any case they are likely to be edited out of books); but we do know from everyday experience that we all make mistakes quite often. A widely quoted figure for speech errors is once every thousand words (Dell 1995), so errors are bound to figure in any corpus of unedited spoken language.

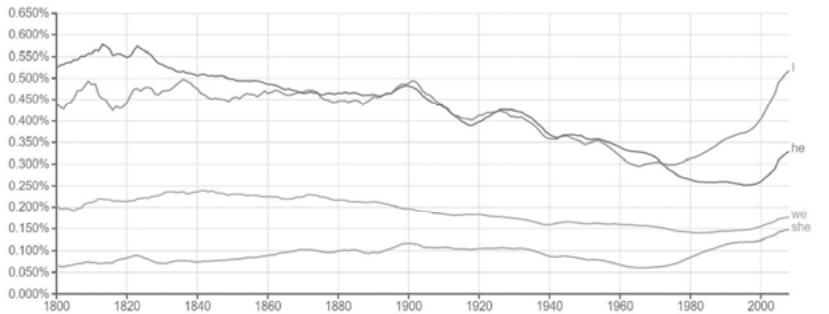


Fig. 6: Changes in pronoun use

Fig. 7: Figures for three common misspellings

Another problem is restricted to written language, where standardised norms are much more clearly codified (e.g. in dictionaries). Writers have to learn these norms, so in addition to typos, written corpora may show the effects of ignorance, such as spelling mistakes. Fig. 7 shows three common misspellings in

Google books: *acheive, apparantly* and *dissapear*. This graph shows that these spellings occur significantly often, but of course it doesn't prove that these spellings are part of the target cognitive system of a competent speaker of English. Presumably the recent reduction in all three errors is due to automated spell-checking.

The conclusion, then, is that a corpus is not an accurate guide to the underlying language system either of the individual language producer (because of errors in speaking and writing) or of the community (because of ignorance in matters such as spelling). On the other hand, the corpus is an accurate reflection of the E-language – Chomsky's 'external language' which he contrasts with I-language or 'internal language'. This being so, a corpus can be taken, with one important reservation, as a description of the language experience of a child learning the language concerned. What a child hears certainly contains speech errors, and may contain errors of ignorance (e.g. if the speaker concerned is another child or a second-language speaker of the language). The challenge for the child is to make sense of this input by building a language system to explain it and enable similar output.

The important, but rather obvious, caveat is that any corpus is a very impoverished account of the child's language experience. For one thing, the child's experience includes the total context, and not just the words, so the child can assign meanings and social significance to the words in a way that is not possible in a disembodied corpus. And for another, the child's experience is gradual and spread over many years, whereas all the data in a corpus is available simultaneously. The child builds I-language gradually, a bit at a time, and interprets new E-language in terms of the existing I-language; so at any given time a lot of E-language is simply incomprehensible, and has no effect on the I-language.

The third link between corpora and cognition is in language processing, which is where a great deal of the recent research on dependency distance is focused (Liu, Xu & Liang 2017; Temperley & Gildea 2018). For this book, this is the most important of the three links. Once again, the link is complex because most of the issues in processing are framed in terms of the needs of a reader or hearer, but readers don't feature in a corpus. What a written corpus shows is the decisions made by writers, so it reflects the thought processes of writing rather than reading; on the other hand, any commercial author pays attention to the reader's needs and wants, so a corpus does reflect these needs as perceived by the writers. However, what we cannot assume is that every sentence or word in a corpus is equally easy to read, and indeed we know that some authors deliberately aim to impress rather than to communicate easily. The criteria for 'good

writing' vary from time to time as part of literary culture, as can be seen in Fig.8. This shows the rise and fall of the phrases *elevated style* and *noble style*. These terms evoke a challenging style characterized by long sentences and complicated syntax, but when used they presumably show admiration for such style. These terms dominated the 19th century, but have since been replaced, and dwarfed, by the phrase *easy reading*, which was later replaced by the more easily-read *easy read*.

Fig. 8: Changing attitudes to style

What these figures show is that attitude to written style is an important influence on writers' syntactic choices. An elevated or noble style favours complex syntax over easy reading; indeed, nobility may imply hard work on the reader's part, so the harder it is to read, the more noble the style. Although these terms are no longer fashionable, we can't take it for granted that modern writers are giving top priority to the reader's processing needs; indeed, some academic papers give the impression that their writers ignore the reader's needs altogether. No wonder that the n-gram for the word *readability* (not shown here) shows a spectacular rise peaking about 1980, presumably in reaction to the prevalence of opaque style.

In short, it's important to be cautious in using corpora as evidence for cognitive processing unless we also know how readable the material is.

3 N-grams, dependency structure or phrase structure

The previous section took it for granted that a corpus can be used as a source of evidence for some kind of investigation, whether cultural, linguistic or psycholinguistic, but of course this need not be so: it can be used simply as a tool for some engineering project such as automatic translation. The wonderful Google N-gram tool is a by-product of this engineering approach. However, for all its utility as a tool, it is a poor basis for understanding how language works because it is tied too closely to the corpus; for instance, if you apply it to the adjacent pairs of words in a bizarre but perfectly grammatical sentence like *Does Dizzy ever bellow?*, it fails to find any 2-word n-grams at all because the adjacent pairs occur so rarely even in the massive Google corpus.

Another objection to n-grams is, of course, that they aren't a plausible structure for syntax – meaning that they aren't cognitively plausible as an explanation for how we deal with sentences. N-grams are simply a statistical description of the input to language learning; but the goal of analysis, whether by the linguist or by the language-learner, is to understand the system underlying this input.

It's true that when we learn language we learn vast amounts of 'formulaic language' (Wray 2000; Wray 2006) – multi-word 'chunks' such as *what I mean is or at the end of or if I were you*. But all these chunks have an internal structure which allows them to integrate gracefully into a sentence. There is no doubt that sentences have structure beyond simple strings of words; the question is what kind of structure sentences have. Answering this question is essential for quantitative syntax because it decides what is available for counting.

As we all know, there are many different answers to this question, differing in a number of different ways. We shall consider three of these differences. One concerns the choice between dependency structure and phrase structure. Fig. 3 shows very clearly that, as a matter of fact, the dominant answer is phrase structure (especially if we include hits for its synonym, *constituent structure*), so it's important for those of us who disagree to be clear about why we reject this answer. The fact that computational linguistics (including corpus studies) takes dependency structure much more seriously than the mainstream of theoretical and descriptive linguistics is not in itself a good argument for preferring it to phrase structure, though it may point to good arguments.

What, then, is the difference between dependency structure and phrase structure? There are two relevant differences:

1) word-word dependencies
2) extra nodes for sequences of more than one word, plus links from these nodes to their parts.

For instance, in the sentence *Small babies cry* we can recognise the structures shown in Fig. 9:

1) word-word dependencies between the pairs <small, babies> and <babies, cry>, represented by curved arrows.
2) two extra nodes: [small babies] and [small babies cry], together with relevant part-whole relations represented by straight lines.

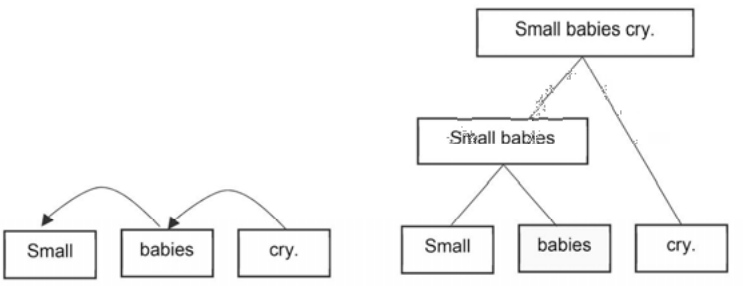


Fig. 9: Dependencies or part-whole relations?

The former is what we call dependency structure while the latter is phrase structure or constituent structure. Logically, of course, a third possibility is to combine the two systems, which has occasionally been suggested (Hudson 1976; Matthews 1981), but for present purposes we can ignore this possibility. The theory of dependency structure was first developed by Sibawayh, an 8th-century Persian grammarian of Arabic (Owens 1988; Percival 1990), but it was first applied in modern times by the German Franz Kern (Kern 1884), followed later by Lucien Tesnière (Tesnière 1959; Tesnière 2015). In contrast, phrase structure was first suggested by another German, the psychologist Wilhelm Wundt (Wundt 1900), and adopted by the American linguist Leonard Bloomfield (Bloomfield 1933; Percival 1976) whose followers included Zellig Harris, tutor of Noam Chomsky. There has been shamefully little debate about the relative merits of the two systems, so their distribution follows social rather than intellectual parameters: dependency structure tends to be used by Europeans and computational linguists, while phrase structure is strongly linked with the USA and American influence. As we have already seen, phrase structure dominates in

linguistics: 'Phrase structure grammar has such a great weight of authority behind it that many linguists take it virtually for granted.' (Matthews 2007:127)

Why, then, should a student of corpora prefer dependency structure to phrase structure? If the aim of such research is to count cognitively real structures, this is a question about truth, not convenience. A dependency structure is obviously simpler because it has no extra nodes for larger units and therefore contains fewer nodes, but that in itself isn't relevant because those extra nodes might be cognitively real. So how can we decide which kind of structure is nearer to the cognitive truth?

The best argument for dependency structure is very simple: why not? The point of this question is that phrase structure denies the existence, or possibility, of relations between individual words. Returning to the earlier example, phrase structure claims that a direct link between the two words is impossible for the human brain to handle, so the only way to show that *small* is related to *babies* is to create an extra node that contains both these words. As a claim about human cognition, this is absurd because we know that we can recognise all sorts of direct relations between individual entities. The most obvious area to find such relations is in our social networks, which must be represented cognitively as a network of individuals with direct relations such as 'brother of' or 'friend of'. When we learn that Kim's mother is Pat, we don't have to create an extra node for the pair of them in order to show this relationship; indeed, this would be a grossly inefficient way to handle it because it would require not only an extra node but also two 'part of' relations, and even then it wouldn't show that Pat is Kim's mother. If, then, we can know that Pat is the mother of Kim, why can't we also know that the word *small* is the dependent (and adjunct) of *babies*? And of course if word-word dependencies really are possible after all, then all the arguments for extra nodes, and phrase structure, need to be reconsidered.

Those who use phrase structure never consider this argument. Their argument is based on the unargued assumption that the only possible kind of structure is phrase structure, so if a sentence is more than an unstructured string of words, it must have a phrase structure. This assumption creates serious problems of its own for grammar-writers:

1) How to express word-word relations such as that between a verb and the preposition that it selects (*depend* selects *on*, *look* selects *after*, and so on), if the only formal route from one to the other lies through at least two extra nodes (one for the VP which holds them together, and another for the PP headed by the preposition).

2) How to explain why so many syntactic patterns can, in fact, be defined in terms of a head word with its dependents. In phrase structure this has to be stipulated as an unexplained fact, but dependency structure builds it into the foundational assumptions. There is some debate as to whether every construction does in fact have a head-dependent structure; for example, it has been argued that *Student after student came into the room* cannot be given such a structure (Jackendoff 2008), but this conclusion is premature given the obvious dependency analysis of examples like *student after student* with the first *student* acting as head and carrying special properties due to its dependent. The assumption behind dependency structure is that every construction has a dependency structure, and the main challenge is as always coordination, where the conjuncts are syntactically equal. It is important to remember that coordination is just as much of a challenge for phrase structure, given that it creates larger units which cut across phrases (as in *I went to Birmingham on Tuesday and Edinburgh on Wednesday*).
3) How to explain word-order tendencies in head-initial and head-final languages, given that these tendencies affect all dependents. In the absence of the notion or term 'dependent', how can we even formulate the generalisation that (say) all dependents follow their head (as in a typical VSO language)? Admittedly some phrase-structure theories recognise a 'head parameter' which is meant to cover such generalisations, but the informal statement conflicts with the assumed theory.

The problems created by assuming phrase structure, but avoided by dependency structure, are significant, but advocates of phrase structure generally ignore them. Nevertheless, they happily talk informally about dependencies when discussing 'long-distance dependencies' (which are syntactic dependencies) and 'anaphoric dependencies' and other 'local dependencies' such as NP movement (which are not) (Manzini 1990); and, of course, since the advent of X-bar syntax, the notion of heads and dependents has been central even to phrase structure (Jackendoff 1977).

The choice between dependency structure and phrase structure is really important when studying corpora quantitatively, because it determines what is available for counting. Both kinds of analysis show individual words and their various features, including their order, but if you want to count dependencies or measure them, you have to use dependency analysis. Equally, of course, you need phrase analysis if phrases are what you want to count; but since there's no

reason to believe that extra phrase nodes are psychologically real, why would anyone want to count them?

As an easy example of what can be achieved with dependency analysis, we can use Google N-grams to explore the two alternative orders after a phrasal verb such as GIVE UP, meaning 'abandon' or 'stop'. As far as the grammar of English is concerned, both orders of *up* and the direct object (Obj + Part or Part + Obj) are generally agreed to be possible:

(1) He gave smoking up.
(2) He gave up smoking.

But if the object is a personal pronoun, it must stand before *up* (giving Obj + Part):

(3) He gave it up.
(4) *He gave up it.

However, it is also agreed that the order Part + Obj is increasingly likely as the object gets longer (Biber et al. 1999:932), which many of us would explain in terms of the length of the dependency between *give* and *up*: speakers and writers try to minimize dependency length.

However, according to Google N-grams there are other influences on the choice of word order. Consider first the figure for BRING UP with *this matter* as object. What emerges from Fig. 10 is that once this combination became established in the vocabulary of English, it followed a regular pattern in which the two orders had roughly the same probability, but with Obj + Part consistently outnumbering Part + Obj.

We can now contrast this pattern with that for another four-word combination: BRING OUT + *this point*. According to Fig. 11, this combination entered English at about the same time as BRING UP + *this matter*, but it has a very different history. Once again, one of the options consistently outnumbers the other, but this time the winner is Part + Obj; and the two options follow very different trajectories due to the very sharp rise and fall of the Part + Obj option. (The other option follows much the same trajectory as for the other combination, BRING UP + *this matter*, so it is clearly the rise-fall trajectory that needs an explanation.)

Fig. 10: BRING UP + object

Fig. 11: BRING OUT + this point

I have no idea how to explain this eccentric pattern, but it's clearly not a matter of dependency length because the dependencies have the same length in both combinations. It's possible that the explanation lies in the relative 'entrenchment' of the different multi-word chunks mentioned earlier (Bybee 2006; Langacker 2007). If so, the figures would be taken to show that the chunk BRING OUT THIS POINT underwent a rise and fall in its entrenchment which it didn't share with its synonym BRING THIS POINT OUT: the more often it was used, the more entrenched it became in people's store of chunks and the more likely they were to use it themselves. Such an explanation would be completely compatible with a cognitive approach to dependency structure, and would indeed bring quantitative dependency studies together with the research field of 'usage-based' studies of constructions (Barlow & Kemmer 2000; Goldberg 2006). Unfortunately, the American roots of this research mean that it almost always assumes a version of phrase structure.

4 Poor or rich structure

The second of the three parameters mentioned earlier on which syntactic analyses differ is in terms of 'richness': how much information do they give? The history of phrase structure is dominated by attempts to enrich the very impoverished structures of the early days. One of the earliest attempts was Eugene Nida's addition of dependency information to distinguish heads and dependents (Nida 1943), as illustrated in Fig. 12 (from page x of Nida's thesis which was eventually published in 1960 and is now available for download at www.sil.org). The horizontal lines are extra phrasal nodes, so that *he* combines not with *ran* but with *ran away*; but the line also carries a symbol which shows the internal structure of these phrases. The arrow head points to the head of an endocentric construction, while the 'X' indicates an exocentric construction without a head – following the ancient tradition of grammatical analysis in which the subject-predicate relation is excluded from the analyses in terms of dependency.

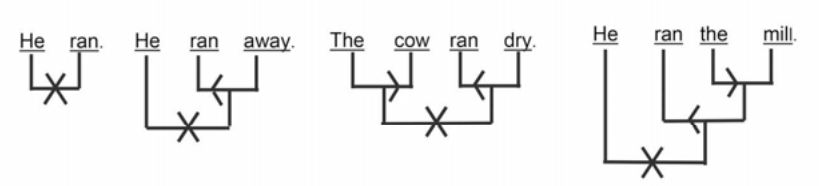


Fig. 12: Nida's enriched phrase structure

Nida's enrichment didn't catch on, but Chomsky's addition of transformations certainly did. Like Nida, Chomsky saw that phrase structure needs some kind of enrichment, but his suggestion was a series of different phrase structures linked by transformation. More recently, Bresnan proposed a two-level analysis in which a phrase structure is combined with a functional structure (Bresnan 2001), and Sag and Pollard, following Gazdar, proposed a single-level analysis in which a phrase structure is enriched with elaborate feature structures (Gazdar et al. 1985; Pollard & Sag 1994).

Meanwhile, very similar theoretical issues arise for dependency structure, suggesting that a simple, unenriched dependency structure is too simple. For example, most linguists in the phrase-structure tradition accept the need for double-motherhood – a structure in which one element belongs to two separate phrases. A typical example, and argument, can be based on example (5).

(5) It was raining.

The proposed analysis recognises that *it* is the subject not only of *was* but also of *raining*, so it must be part of a discontinuous phrase *it raining*. The link to *was* is obvious and uncontentious, but that to *raining* is equally clear, because the verb RAIN requires *it* as its subject. This is a purely syntactic rule because *it* has no referent, so it's not required by the meaning; but it is strict, so it applies, for instance, even if RAIN is a gerund. In examples (6) and (7), the notation **(it)* means 'is obligatory, cannot be omitted'.

(6) I remember *(it) raining.
(7) *(It) raining spoilt the party.

In short, the verb RAIN, really does need a syntactic subject, *it*, and this need applies just as strongly when the verb is non-finite, so *it* must be the subject of *raining* in example (5). The evidence shows, therefore, that one word may depend (in this case, as subject) on more than one other word.

Phenomena such as multiple dependencies show that simple dependency structures need to be enriched in some way. As in the phrase-structure world, one suggestion for enrichment is to combine a number of different simple structures, each representing one of several levels between meaning and phonology (Mel'čuk 2003); in such an analysis, *it* would depend on *was* on one level, and on *raining* on another. One problem for such an approach is that this pattern is indefinitely recursive, as in example (8), which seems to require an indefinite number of intervening levels.

(8) It seems to have been about to start raining.

Another suggestion is to combine a simple tree with the valency frames that license it. Verbs such as *was* would have a valency frame that shows the double dependency of the subject, so this information need not be included in the sentence structure (Osborne & Reeve 2017). The objection is obvious: a syntactic analysis reflects the valency frames from which it inherits, so it should include all the information that it inherits.

My own preferred option is to abandon the simple view of dependency structures and accept that they need the richness of networks (Hudson 2007; Cong & Liu 2014). From a cognitive viewpoint, we know that the human mind is capable of accommodating networks – we all accommodate a vast social network, and a common view in cognitive science is that the whole of knowledge

takes the form of a network (Reisberg 2007). It's only simple-minded linguists who benefit from the simplicities of a simple tree-like structure for syntax. In a network, any node may be linked to any number of other nodes, so multiple dependencies are easily accommodated. In this approach, the structure for our example is the one shown in Fig. 13.

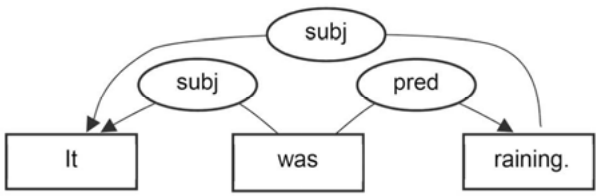


Fig. 13: A network structure for *It was raining*.

The network principle allows us to accommodate complexity where it arises, but it should be emphasised that complexity is the exception rather than the rule. The structure assigned is exactly as complex as the sentence requires, so most sentences have simple tree-like dependency structures like the one in Fig. 14. (The function labels are abbreviated as follows: adjunct, complement, subject.)

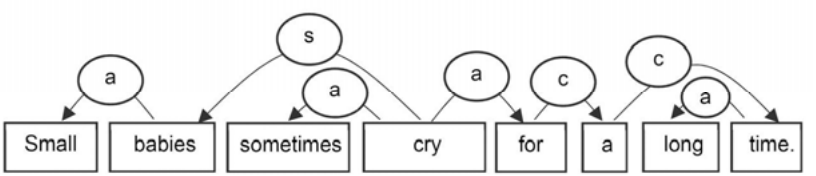


Fig. 14: A simple dependency structure for a long sentence

At the other extreme, some sentences need a complex dependency structure such as the one in Fig. 15 (where o stands for 'object', x for 'extractee' and *p* for 'predicate').

The relation between *what* and *do* in this structure is particularly important, because it is a case of mutual dependency: *what* depends on *do* as its extractee (i.e. *what* is 'extracted' from its normal position and positioned before *do*), but *do* depends on *what* because *what* takes a finite verb as its complement. Mutual

dependency is absolutely impossible to present in a tree-like structure where the vertical dimension shows subordination, because *what* cannot be simultaneously above and below *do*.

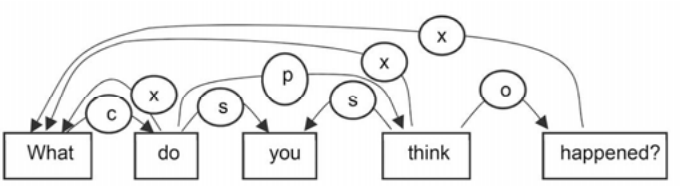


Fig. 15: A complex structure for a complex sentence

Every link in this structure can be justified by ordinary linguistic arguments which guide us to a cognitively accurate structure. (Indeed, I have argued elsewhere (Hudson 2018) that the structure actually needs a number of extra nodes for distinct tokens of the same word, because the whole point of adding a dependent is to modify the head word; so we need a token node for unmodified *do*, but we also need one for *do* as modified by the subject *you* and another for it when modified by the predicate *think*.) Consequently, every link is part of the sentence's cognitively real structure and should be included in any analysis which claims to be 'complete'.

However, this is obviously not to say that every research project on quantitative syntax must include every link in its database. It all depends on the purpose of the project. For example, if the aim is a valency dictionary, then it's important to include all subjects and complement links, even if they're shared; so the analysis must show that, in *It was raining*, *it* is the subject of *raining* as well as of *was*. Otherwise this would be a case of *raining* without the subject *it*, which is in fact grammatically impossible (as I showed earlier). On the other hand, it would be right to ignore the relation between *it* and *raining* if the aim is to measure dependency distance, because this relation is predictable from the previous structure – i.e. having heard *it was*, the hearer can predict that *it* will be the subject of any predicate that follows.

Returning to Google N-grams, the case of *it was raining* provides an interesting example of how different influences interact in corpus statistics. Fig. 16 shows the figures for both *it rained* and *it was raining* (in this case with case-sensitivity turned off). The point of including this diagram is that it shows how the probability of the lexeme RAIN interacts both with the probability of the accompanying constructions and with the things talked about. I assume that

the gigantic increase in *it rained* between 1810 and 1820 corresponded to some meteorological event (possibly the 'year without a summer' in 1816, due to the eruption of Tambora). But it's noticeable that this has no apparent counterpart in *it was raining*, and also that, compared with other verbs, the progressive equivalent was surprisingly rare around 1800. On the other hand, the two sentences do show similar trajectories in the second half of the 20th century, suggesting that by that time the effect of the lexeme (and the weather) was outweighing that of the progressive.

Fig. 16: It rained and It was raining.

5 Universal or parochial structure

A third parameter on which syntactic analyses vary is the extent to which they apply the same analysis, with the same categories and the same dependencies, to every language. This is the contrast between a universal grammar and a 'parochial' grammar designed for just one language. This contrast has divided linguists for centuries; for example, in the early 20th century British grammarians (more accurately, writers of school grammars) were deeply divided between the followers of Edward Sonnenschein, Professor of Classics at Birmingham University, taking a strong universal line, and Otto Jespersen, arguing that every language should be described in its own terms (Walmsley 1989).

If the aim of the analysis is cognitive reality, this dispute amounts to a theoretical debate about whether or not there is a cognitively real and innate Universal Grammar. This isn't the place for a discussion of this question, but the onus is on those who support Universal Grammar to provide strong evidence. After all, it's obvious that grammar varies from language to language, and it's

equally obvious that children learn the grammar of their language on the basis of usage – what they hear other people saying – so the null hypothesis must be that children learn their entire grammar from usage, without any input from innate Universal Grammar. In my opinion, there is no compelling evidence against the null hypothesis; on the contrary, I believe that the challenge is to explain not the similarities among languages, but their differences (Evans & Levinson 2009).

If this is right, then quantitative corpus studies across languages face a serious problem of comparability. How can we compare the syntactic structures of corpora from two different languages? To illustrate this problem, take Latin and English. Latin has case inflections which are often translated in English as prepositions. For example, the dative of DOMINUS, 'master' is *domino*, which might be translated into *to the master*. For centuries, the universalists have hidden this difference by pretending that in some sense *to the master* is the dative of MASTER, in just the same way that they treat *will go* as the future tense of GO. For an example of this tradition we can return to the German inventor of dependency structure, Franz Kern, (described elsewhere as 'the father of the syntactic tree' (Liu 2009:5)) who gives the diagram in Fig. 17 (redrawn here for clarity) as an example (Kern 1884:30). The German sentence in the example is (9).

(9) Eine stolze Krähe schmückte sich mit den ausgefallenen
a proud crow decorated himself with the fallen-out
Federn der Pfauen.
feathers of the peacocks.
"a proud crow decorated himself with the feathers dropped by the peacocks."

This example is, to the best of my knowledge, the first published example of a complete dependency analysis with (crucially) the verb as the sentence root, and therefore predates Tesnière by some decades; but for all its innovation in this respect, in other respects it is highly conservative. In particular, it treats the preposition *mit*, 'with', as though it was a case inflection – in spite of the fact that *mit* is not even next to the supposedly case-bearing noun *Federn* (and that this noun really is inflected for case, but with the dative case demanded by *mit*). The motivation for this analysis is clearly to bring German into line with Latin, where the meaning 'with' might have been expressed simply by an ablative case without a preposition. This is the kind of analysis that Jespersen argued so strongly against.

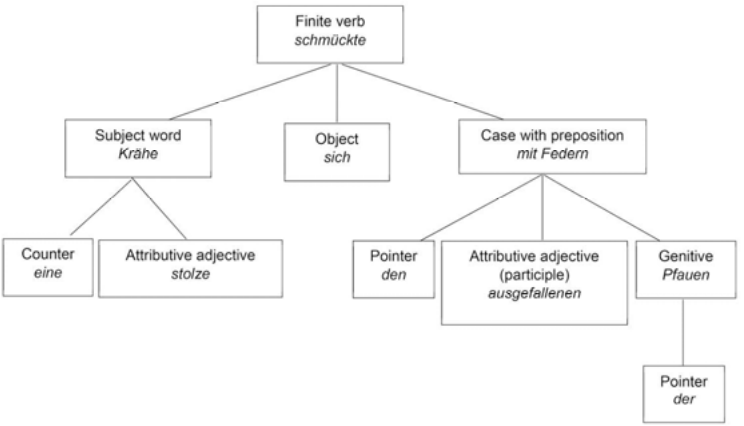


Fig. 17: German dependency structure by Kern

To show that this issue is still with us, and still unresolved, consider the Stanford parser, which

> 'provides a uniform analysis of prepositions and case in morphologically rich languages. The analysis we chose is to push all the way the design principle of having direct links between content words. We abandon treating a preposition as a mediator between a modified word and its object, and, instead, any case-marking element (including prepositions, postpositions, and clitic case markers) will be treated as a dependent of the noun it attaches to or introduces' (de Marneffe et al. 2014).

One immediate benefit of this approach is that languages are easy to compare, since the same set of universal categories (at least for case) is used in analysing corpora from every language. But at what cost?

This is the same principle as the one which brought grammar into disrepute centuries ago by forcing English into the mould of Latin: since Latin had an ablative case, English must have one too, but marked by preposition rather than by morphology. What if this analysis is actually wrong for English? This question leads to the frontiers of research in theoretical and descriptive linguistics, and deserves serious debate. Minimalism claims that English has a rich case system (Chomsky 1995:110), but some of us have argued that it has no case at all (Hudson 1995). It's not just a matter of taste or convenience; once again, it is ultimately a matter of cognitive truth. There is relevant evidence, such as the strong tendency for pronoun forms (e.g. *me* or *I*) to be influenced by the presence of coordination as in (11) and (13):

(10) You and I are friends.
(11) You and me are friends.
(12) It's for you and me.
(13) It's for you and I.

This is a change in progress, but it has been in progress for a long time and shows up even in the edited books in the Google N-gram collection. Fig. 18 shows a gradual rise in the coordination-influenced uses (as well as some mysterious but synchronised fluctuations in their alternatives). It is tempting to dismiss patterns like *you and me are* as simply wrong, but there must be some motivation for them, and some reason why German, which really does have case, shows no such change. One possible explanation is that English no longer has any case, and the choice of pronoun forms, the last relic of our old case system, is determined by local rules which apply only to these forms and which (unlike their case-based ancestors) pay attention to coordination as well as to grammatical function.

Fig. 18: Pronoun form and coordination

The point of this discussion is to illustrate the tension between comparability and truth. Forcing all languages into a single framework of analytical categories, such as a system of cases, makes them easy to compare; but the comparison will be meaningless if the analysis is wrong: garbage in, garbage out. To take a simple example, consider the English phrase *went to Rome* and its Latin translation *Romam ivit*, and the simple question: how many dependencies link adjacent words? If *to* is a preposition linking *went* to *Rome*, then the English phrase has two dependencies and both are between adjacent words; but if *to* is a mere case-marker, it separates *Rome* from the word on which it depends: *went*.

Since the example illustrates a very common pattern, the syntactic analysis chosen will have a major effect on the statistics of a corpus, giving either a relatively high adjacency count for English or a relatively low one. And of course at least one of these counts is wrong, or at least meaningless or misleading.

Where does this leave language comparison? Is it possible to compare languages if the categories used in analysing their sentences are different? I believe that comparison can in fact be reconciled with truth, though it requires caution. There are three reasons for this belief:

1) the universality of dependency structure
2) similar functional pressures
3) bilingual code-mixing.

The first reason is that the basis for any syntactic theory is the assumption that every language has syntax, which means that words can be combined so as to express complex meanings. Combinatorial meaning leads inevitably to subordination, in which one word provides a general meaning which a dependent word makes more precise: so a big book is a particular kind of book, and John snoring is a particular kind of snoring. Consequently, we can assume, as a matter of general theory, that every language has dependency structure, so the notion 'dependency' is a universal, and dependency structures can indeed be compared across languages even if we have to ignore the labels on the words and links.

The second reason for believing that comparisons can be true is that grammars are shaped by functional pressures, and that, since pressures conflict, every language produces its own 'engineering solution' (Evans & Levinson 2009) which may distinguish it from other languages. On the other hand, the main functional pressures define a limited space for variation, so different languages may, in fact, end up with very similar solutions (and especially so if there is some contact between their speakers). For example, languages typically distinguish nouns and verbs, and if a language has one-place verbs such as 'sleep' and 'run', it allows these verbs to combine somehow with a single dependent noun which may then be aligned with either of the dependent nouns of a two-place predicate – hence the contrast between so-called nominative-accusative languages and ergative languages. Either way, we can call the one-place dependent the 'subject', which allows us to compare the subjects of this language with those of any other language.

What we must avoid, however, is the assumption that translation equivalents have the same syntactic structure. This temptation is responsible not only for 'Latinate' grammars of English (which forbid *to boldly go* because *to go* cor-

responds to a single word in Latin), but also for 'Englishate' grammars of other languages which assume a universal underlying structure which is basically the structure of English. A case in point is the analysis of determiners. A very strong case can be made for the so-called 'DP analysis' (Determiner Phrase) of English in which determiners are heads, so that in *these books* the common noun *books* depends on *these* rather than the other way round (Hudson 1984; Abney 1987). But that doesn't mean that the same analysis can be applied automatically to the translation equivalents of English determiners in every other language. In some languages, the evidence is even more compelling; for instance, in French and many other Western-European languages, prepositions regularly fuse with a following definite article (e.g. the expected *de le* meaning 'of the' is written *du*), showing that the article must depend directly on the preposition. But in other languages there is very little evidence even for the existence of a category 'determiner', let alone for the DP analysis. For instance, Latin nouns can occur freely on their own, so *dominus* can mean either 'a master' or 'the master', which removes the main evidence for English determiners. Consequently, the traditional classification as adjectives of Latin words with meanings such as 'this' or 'every' may be correct, and Latin may have no determiners.

My third reason for believing that languages can be compared comes from bilingual code-mixing (Wang & Liu, 2013; Wang & Liu, 2016). Bilingual speakers often mix their languages when speaking to other bilinguals. Here are two examples from a study of German-English bilingual speakers (Eppler 2011:9, 18):

(14) Zwei Tage spaeter wurde er executed.
two days later was he executed.
"Two days later he was executed."

(15) You don't need to wegwerfen.
You don't need to throw [it] away.

What such examples suggest is that bilingual speakers make their languages interchangeable by using the same categories for both languages. In example (14), the German *wurde* requires a past participle, but not necessarily a German past participle, so an English past participle will do as well as a German one. Example (15) illustrates the converse relation, where an English word *to* has a need for an infinitive which is satisfied by a German infinitive (marked in German by the suffix {en}). We can therefore conclude that the categories 'past participle' and 'infinitive' apply equally to both languages, and must be psychologically real for at least these speakers.

But what about case, which certainly applies to German but (arguably) not to English? Here are three relevant examples (Eppler 2011:32, 44, 196, 224):

(16) die Hungarians, die Czechs, die haben immer
the Hungarians the Czechs they have always
a worse accent than we have.
a worse accent than we have.

(17) ja er muss ins Spital fuer diese prostate gland operation.
well he must into.the hospital for this prostate gland operation.
"Well, he has to go into hospital for this prostate gland operation."

(18) wegen einer Bombe – a bomb scare
because.of a bomb a bomb scare

(19) eat it with der Hand – because das schmeckt ganz anders.
eat it with the hand because that tastes quite different.
"Eat it by hand because that tastes quite different."

The first two examples have English words or phrases embedded in a German context where a particular case is required: nominative in (16), accusative (required by the preposition *fuer*) in (17). These examples could be used as evidence that English nouns do indeed have case, even though the cases aren't morphologically distinguished; but a much easier explanation is that case restrictions only apply to words that have case, namely German nouns, pronouns and adjectives, so they can be ignored in English nouns. This explanation also avoids the problem of the genitive, which for advocates of English case is expressed morphologically by {s} as in *the man's hat*. Example (18) challenges this analysis because the German preposition *wegen* requires a genitive, and receives a German genitive *einer* (marked morphologically by {er}), but then when the noun is changed into English it appears without the predicted {s} of the supposed English genitive.

Example (19) is particularly interesting because of the complex interplay between the two grammars: the English preposition *with* does not govern a case, but its German equivalent, *mit*, does. Accordingly, when the language changes into German, we find the dative case (marked by {er} in *der*) required by *mit*. Cognitively, this is what we expect: a German article always has a case, and has no case-neutral form, so if a German article is used, its case must be selected by the syntactic context. Since the English syntactic context does not require case, the speaker constructs an imaginary German context that does, and thus solves the problem.

If these three arguments are correct, it is possible to compare languages while remaining faithful to their individual grammars. There is no need to force them into a supposedly universal analytical framework which ignores the grammatical evidence and which therefore distorts the results of any comparison. However, there is a significant cost: every language must first be analysed in its own terms, just as Jespersen argued; and only then will it be ready for comparison.

6 Projectivity and word order

Finally, we come to the matter of word order. One of the attractions of dependency structure compared with phrase structure has always been its flexibility in dealing with word order, and especially so when compared with the binary versions of phrase structure. Indeed, this is probably one of the reasons why dependency structures are popular (Liu 2009 2010) where word order is relatively free, and why phrase structure has flourished in the world of English, where word order is rather rigid. It could even be suggested (though I would dispute it) that some languages have dependency structure whereas others use phrase structure; something like this suggestion seems to lie behind the typological distinction between configurational and non-configurational languages (Pensalfini 2006).

Historically, dependency structures have tended to be tied more closely to meaning than to word order; the earliest approaches to dependency, by Panini (Shukla 2006), were purely semantic, and pure syntax has had fluctuating fortunes over the centuries since then: powerful in 8th century Arabic grammar, but until recently much weaker in Europe. Dependency grammarians still disagree about the link between dependencies and word order, but in terms of a single issue: 'projectivity'. This is the question of whether the 'phrases' nest neatly inside one another. In a projective tree structure (such as Kern's diagrams or Tesnière's stemmas), each word 'projects' straight up to the node in a tree structure that represents it without crossing any dependency lines. The lefthand diagram in Fig. 19 is projective because the dotted projection lines don't cross any other lines, but the righthand one is non-projective. Without projectivity (or some equivalent), it is impossible to explain why **Red drink wine!* is not possible, because the individual dependencies follow the same word-order rules in both examples:

1) *wine* is the object of *drink* and therefore follows it.
2) *red* modifies *wine*, and therefore precedes it.

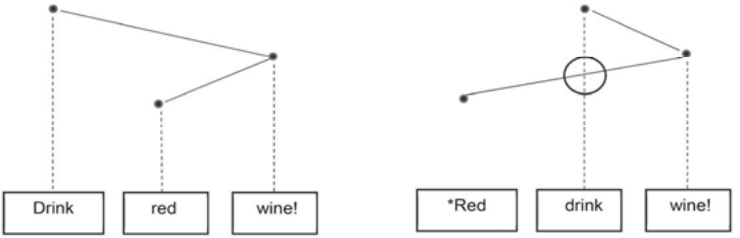


Fig. 19: Projective and non-projective structures

Projectivity is cognitively important because it guides hearers and readers in their search for dependencies. When we read the word *red*, we expect it to depend either on the next word or on some later word; in principle, this could be any distance from *red*, as in *red French freshly-made wine*, but projectivity applies an important limitation on the words to be considered: they mustn't themselves depend on a word that separates them from *red*, as *wine* does in **Red drink wine*. The same principle is of course fundamental to automatic parsing, so it is generally incorporated, in some form, into parsing algorithms.

On the other hand, we have seen strong evidence for rich dependency structures in which words can depend on several other words, which inevitably means that the structures are not projective. Here again, in Fig. 20, is the structure proposed in Fig. 13 for *It was raining*, but converted to stemma format in order to show projectivity violations. Even in this simple example it is easy to see the violation: *it* depends on *raining*, but is separated from this word by a word on which they both depend. If we describe dependencies as relations between a dependent and a head, then *it* has a non-projective head. The much more complex example from Fig. 15 (*What do you think happened?)* would show many more violations, except that it is impossible to present as a stemma because of the mutual dependency.

The challenge, of course, is to explain why both **Red drink wine!* and *It was raining* produce projectivity violations, but one is grammatical and the other is not. This challenge faces computational linguists (Nivre 2006) as well as theoretical linguists. Fortunately, the solution is rather obvious: in the grammatical example, the word that has a non-projective head also has a projective one. That is, *it* depends non-projectively on *raining* but also depends projectively on *was*; whereas in the ungrammatical example *red* only has a non-projective head.

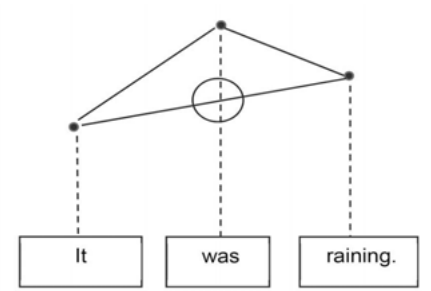


Fig. 20: A grammatical projectivity violation

This contrast has lain behind the various explanations of projectivity offered in Word Grammar since its early days (Hudson 1984:101), all of which assert that every word (except the sentence root) needs one projective head, but may have non-projective heads as well. Moreover, the projective head is typically higher in the dependency hierarchy than the non-projective ones, and therefore it is the projective head that licenses the non-projective heads: 'raising' is permitted, but 'lowering' is typically not allowed. For example, it is part of the valency of *was* to have a predicate complement whose 'raised' subject is also the subject of *was*. Alternatively, the dependent word itself can license the non-projective head, as in an example such as *What do you think he said?*, where *what* first licenses its projective link as 'extractee' to *do*, which then recursively licenses a series of similar (but non-projective) links to the dependents of *do*. In virtually all such examples, the projective dependency comes first, and provides a guide to the later dependencies.

In short, the non-projective dependencies are predictable from the projective ones, so they are very easy for a reader or hearer to identify. This is an important conclusion for quantitative corpus studies because it reconciles truth with practicality: a true syntactic structure can be very complex, but most of the complexity is predictable. This means, on the one hand, that predictable dependencies can be safely ignored in measures of processing difficulty, but on the other hand it also means that they can easily be added during automatic parsing. For processing difficulty, *It was raining* is no harder than *He likes chocolate*; but its structure includes a (predictable and non-projective) dependency between *it* and *raining* which could be found in a lexical search for dependents.

7 Conclusion

The main conclusion is that dependency structure is better than phrase structure as a tool for representing the syntax of a corpus because it is cognitively – i.e. psychologically – more realistic. The underlying assumption is that cognition is relevant to corpus studies, in spite of the traditional separation of the two approaches to language, because it is the ultimate criterion of truth: there's no point in pursuing an analysis that demonstrably conflicts with the analyses that native speakers assign in their minds.

However, this argument does not lead to the simple dependency structures that are generally assumed in corpus analysis. If dependency structures are psychologically true, then they are also complex – networks rather than trees. Some of the links concerned reflect surface matters such as word order, while others are concerned with deeper matters such as meaning, but all are integrated into a single structure. On the other hand, most of the complexity is predictable either from general rules (e.g. for extraction) or from lexical entries for individual words (e.g. for raised subjects), so it can be added quite easily to a parsed corpus. And of course once the extra links are added, it is up to the researcher to decide whether or not to record them in a corpus survey.

The principle of cognitive truth also conflicts with universal analytical frameworks because each language has a unique grammar which allows unique structures. As we all know, we can't assume that translation equivalents in different languages must have similar syntactic structures. Nor can we assume that different languages must recognise the same grammatical categories, such as word classes and functions. However, the variation is limited by the fact that all languages face similar functional pressures, so some comparison is in fact possible. But of course it's a matter for research to determine which features are comparable across which languages.

In short, dependency analysis of corpora can throw considerable light on cognition, provided the analysis is done with due attention to what we already know about cognition.

References

Abney, Steven. 1987. *The English Noun Phrase in Its Sentential Aspect*. MIT.

Barlow, Michael & Suzanne Kemmer. 2000. *Usage Based Models of Language*. Stanford: CSLI.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart and Winston.
Bresnan, Joan. 2001. *Lexical–Functional Syntax*. Oxford: Blackwell.
Bybee, Joan. 2006. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.
Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
Cong, Jin & Haitao Liu. 2014. Approaching human language with complex networks. *Physics of Life Reviews, 11*(4), 598–618.
De Marneffe, Marie–Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC),* 4585–4592.
Dell, Gary. 1995. Speaking and misspeaking. In Lila Gleitman, Daniel Osherson & Mark Liberman (Eds.), *An Invitation to Cognitive Science: Language* (pp.183–208). Cambridge, MA: MIT Press.
Eppler, Eva. 2011. *Emigranto: The Syntax of German–English Code-switching*. (Austrian Studies In English, Band 99. Herausgegeben von Manfred Markus, Herbert Schendl, Sabine Coelsch–Foisner). Vienna: Braumüller.
Evans, Nicholas & Stephen Levinson. 2009. The Myth of Language Universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences, 32*(5), 429–448.
Gazdar, Gerald, Ewan Klein, Geoffrey Pullum & Ivan Sag. 1985. *Generalized Phrase Structure Grammar.* Oxford: Blackwell.
Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
Hudson, Richard. 1976. *Arguments for a Non–transformational Grammar.* Chicago: Chicago University Press.
Hudson, Richard. 1984. *Word Grammar*. Oxford: Blackwell.
Hudson, Richard. 1995. Does English really have case? *Journal of Linguistics, 31*(2), 375–392.
Hudson, Richard. 2007. *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
Hudson, Richard. 2018. Pied piping in cognition. *Journal of Linguistics, 54*(1), *85–138.*
Jackendoff, Ray. 1977. *X–bar Syntax: A Study of Phrase Structure.* Cambridge, MA: MIT Press.
Jackendoff, Ray. 2008. Construction after construction and its theoretical challenges. *Language, 84*(1), 8–28.
Kern, Franz. 1884. *Grundriss der Deutschen Satzlehre.* Berlin: Nicolaische Verlags–Buchhandlung.
Langacker, Ronald. 2007. Cognitive grammar. In Dirk Geeraerts & Hubert Cuyckens (Eds.), *The Oxford Handbook of Cognitive Linguistics* (pp.421–462). Oxford: Oxford University Press.
Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.
Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120*(6), 1567–1578.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193.
Manzini, Rita. 1990. A new formalization for locality theory. In Joan Mascaro & Marina Nespor (Eds.), *Grammar in Progress: GLOW Essays for Henk van Riemsdijk* (pp.323–330). Dordrecht: Foris Publications.
Matthews, Peter. 1981. *Syntax*. Cambridge: Cambridge University Press.

Matthews, Peter. 2007. *Syntactic Relations: A Critical Survey*. Cambridge: Cambridge University Press.
Mel'čuk, Igor. 2003. Levels of dependency in linguistic description: Concepts and problems. In Vilmos Agel, Ludwig Eichinger, Hans–Werner Eroms, Peter Hellwig, Hans Jürgen Heringer & Henning Lobin (Eds.), *Dependency and Valency: An International Handbook of Contemporary Research, Vol. 1* (pp.188–229). Berlin: de Gruyter.
Nida, Eugene. 1943. *A synopsis of English syntax*. UCLA.
Nivre, Joakim. 2006. Constraints on non–projective dependency parsing. Eleventh Conference of the European Chapter of the Association for Computational Linguistics.
Osborne, Timothy & Matthew Reeve. 2017. Control vs. Raising in English: A Dependency Grammar Account. *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 176–186. Pisa.
Owens, Jonathan. 1988. *The Foundations of Grammar: An Introduction to Mediaeval Arabic Grammatical Theory*. Amsterdam: John Benjamins.
Pensalfini, Rob. 2006. Configurationality. In Keith Brown (Ed.), *Encyclopedia of Language & Linguistics* (pp.23–27). Oxford: Elsevier.
Percival, Keith. 1976. On the historical source of immediate constituent analysis. In James McCawley (Ed.), *Notes from the Linguistic Underground* (pp.229–242). London: Academic Press.
Percival, Keith. 1990. Reflections on the History of Dependency Notions in Linguistics. *Historiographia Linguistica, 17*(1), 29–47.
Pollard, Carl & Ivan Sag. 1994. *Head–Driven Phrase Structure Grammar*. Chicago: Chicago University Press.
Reisberg, Daniel. 2007. *Cognition: Exploring the Science of the Mind* (Third media edition). New York: Norton.
Shukla, Shaligram. 2006. Panini. In Keith Brown (Ed.), *Encyclopedia of Language & Linguistics (Second Edition)* (pp.153–155). Oxford: Elsevier.
Temperley, David & Daniel Gildea. 2018. Minimizing Syntactic Dependency Lengths: Typological/Cognitive Universal? *Annual Review of Linguistics*, *4*, 67–80.
Tesnière, Lucien. 1959. *Éléments de syntaxe structurale*. Paris: Klincksieck.
Tesnière, Lucien. 2015. *Elements of Structural Syntax*. (Trans.) Timothy Osborne & Sylvain Kahane. Amsterdam: John Benjamins.
Walmsley, John. 1989. The Sonnenschein v. Jespersen Controversy. In Udo Fries & Martin Heusser (Eds.), *Meaning and Beyond. Ernst Leisi zum 70. Geburtstag* (pp.253–281). Tuebingen: Gunter Narr Verlag.
Wang, Lin & Haitao Liu. 2013. Syntactic variation in Chinese–English code-switching. *Lingua, 123*, 58–73.
Wang, Lin & Haitao Liu. 2016. Syntactic differences of adverbials and attributives in Chinese–English code–switching. *Language Sciences, 55*, 16–35.
Wray, Alison. 2000. Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463–489.
Wray, Alison. 2006. Formulaic Language. In Keith Brown (Ed.), *Encyclopedia of Language & Linguistics (Second Edition)* (pp.590–597). Oxford: Elsevier.
Wundt, Willem. 1900. *Völkerpsychologie, eine Untersuchung der Entwicklungsgesetze von Sprache, Mythus, und Sitte. Erster Band: Die Sprache. Zweiter Teil*. (Second Edition 1904). Leipzig: Engelmann.

Hongxin Zhang, Haitao Liu*

Interrelations among Dependency Tree Widths, Heights and Sentence Lengths

Abstract: This paper chooses two news-genre dependency treebanks, one in Chinese and one in English and examines the synergetics or the interrelations among dependency tree widths, heights and sentence lengths in the framework of dependency grammar. When sentences grow longer, the dependency trees grow both taller and higher. The growths of heights and widths are competing with each other, resulting in a minimization of dependency distance. The product of the widest layer and its width corresponds to the sentence length. These correlations are integrated into a tentative synergetic syntactic model based on the framework of dependency grammar.

Keywords: synergetics; dependency tree; tree width; tree height; sentence length; dependency distance; valency

1 Introduction

Syntax studies the rules/principles and processes governing the structure of sentences. Sentences are ordered in a fashion that is of at least two dimensions (Ziegler & Altmann 2003). Not only is there the linear, horizontal order, which is captured by word sequences, but also a vertical pattern. The vertical layers arise from the grammatical constructs which are nonlinearly ordered in the brain but have to be ordered in the sentences and will be decoded by the reader/listener. Tesnière (1959: 12) claimed that syntactic study in its essence is the transfer between the linear structure and its two-dimensional tree structure. Constructing sentences is to establish the connections (in Tesnière's term, *connexion*) among words, which gives life to a sentence. Conversely, to comprehend a sentence is to understand this kind of *connexion*.

Different from the study of phonetics or morphology, the study of syntax is closely related to the models or frameworks of syntax. Contemporary syntactic studies are based on either dependency grammar or phrase-structure grammar

Hongxin Zhang, Zhejiang University, Hangzhou, P.R.China
Haitao Liu, Zhejiang University, Hangzhou; Guangdong University of Foreign Studies, Guangzhou, P.R.China, htliu@163.com

https://doi.org/10.1515/9783110573565-002

(constituency grammar), or a combination of both. Köhler and Altmann (Köhler & Altmann 2000; Köhler 2012) investigated the interrelations among embedding depths, positions, lengths and complexities of nodes in phrase-structure syntactic trees and hence established the first synergetic syntactic model based on the framework of phrase-structure grammar. Simply put, synergetics examines how elements in a system cooperate (and compete) to make the whole system function (Haken 1978). Providing a framework for the constructions of linguistic theories, synergetic linguistics regards language as a system of self-organisation and self-regulation (Köhler 1986, 2012), " a special kind of dynamic system with particular properties" (Köhler 2012: 169).

But the study of synergetics in the framework of phrase-structure grammar only is not sufficient for the discovery of syntactic/structural regularities. Köhler and Altmann (2009) mentioned that there was a lack of discussion on how relevant variables interact within the framework of dependency grammar. In addition, changing the framework might make it possible to define some new properties and even some new units. Among many others, dependency distance, dependency tree layers, dependency tree heights, dependency tree widths, generalized valencies are some typical units and properties, whose definitions will be given later in this paper.

In recent years, there have been some studies of interrelations between variables in the framework of dependency grammar. For instance, Liu and Jing (2016) examined the interrelations between the number of nodes (or sentence lengths) and the number of layers (or dependency tree heights) in dependency trees. Liu (2017) examined the distribution patterns of layers. Jing and Liu (2015) discovered the interrelations between average dependency distances and the dependency tree heights. Jiang and Liu (2015) examined the influences of sentence lengths on dependency distances. Wang and Liu (2014) found that the most complicated and longest structures tend to occur more at the end of the sentence while the relatively shorter and less complicated structures tend to occur more initially. But these researches still do not suffice to make a circuit of interrelations among variables and are thus unable to constitute a synergetic model thereby.

The research objective of this study is to investigate the structural relationship between one-dimensional linear sentences and their two-dimensional dependency trees (in particular the dependency tree widths and heights) within the framework of dependency grammar, to examine the forces that help shape the dependency trees, and therein to build a synergetic syntactic model among properties/units. To ensure more potential universality of the model, we need to

examine at least two languages different in typology. We choose English and Chinese as the two languages under consideration.

Like other types of trees, layers of dependency trees come as a result of parenthood, which in turn results from governing. The syntactic structure is determined by the relation between a word (a head) and its dependents located below it, thus generating layers. We will follow the operation in Liu (2017) and define the root (usually a verb) of the sentence as located at Layer 1. The dependency tree height is the number of layers of the dependency tree. The widest layers are those layers bearing the most nodes in a tree. For simplicity, in rare cases where there is more than one such layer bearing a maximum number of nodes, we only take the first such layer as the widest one. The dependency tree width is defined as the number of nodes at the widest layer of a dependency tree.

Following the previous definitions, Fig. 1, an exemplary dependency tree bears a height of 5 with 5 layers. The widest layer is Layer 3 with 5 nodes; thus we claim the tree bears a width of 5.

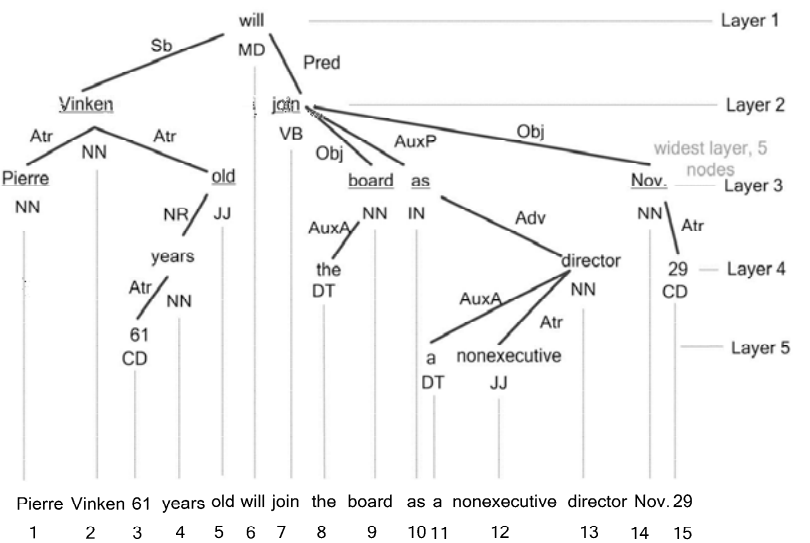


Fig. 1: Presenting a syntactic tree in the framework of dependency grammar

It's interesting to observe that in Fig. 1, the product of the widest layer (3) and its width (5) is exactly the sentence length (15). Will this happen to other sentences?

Therefore, we pose the following research questions.

Question 1: When sentences grow longer, the dependency trees grow taller, generating more layers. How do we quantitatively measure the interrelation between the sentence lengths and numbers of layers (or dependency tree heights)?

Question 2: Similarly, when sentences grow longer, the dependency trees grow wider. How do we quantitatively measure the interrelation between sentence lengths and corresponding tree widths?

Question 3: What is the interrelation between dependency tree heights and widths?

Question 4: What are the relations among the sentence length, the widest layer of the dependency tree, and the number of nodes at this layer?

Fig. 2 visualizes the four research questions. Whether a synergetic model could be built relies on the answers to the first three research questions since the interrelations of the three variables might constitute a circuit.

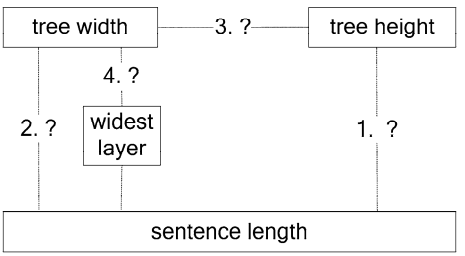


Fig. 2: A synergetic model to be built

In the remainder of this paper, the next section elaborates on the research method and materials. Section 3 details the research findings alongside their analysis. The final section presents conclusions, limitations of the study, and proposals for further research endeavours.

2 Research methods and materials

We choose as our research materials two corpora or syntax treebanks of the news genre, both manually parsed with dependency relations. One is the English part of the Prague Czech-English Dependency Treebank 2.0 (PCEDT 2.0)

(Hajič et al. 2012) and the other is PMT 1.0 (Peking University Multi-view Chinese Treebank 1.0) (Qiu et al. 2014). As both treebanks are annotated with quite similar schemes, a cross-language study is rendered possible. Containing the entire *Wall Street Journal* (*WSJ*) Section of Penn Treebank from the Linguistic Data Consortium (Marcus et al. 1999), PCEDT 2.0 has 1.174 million words with 2,499 stories. PMT 1.0 goes with a size of 3,360,000 words, or 14,463 sentences, including all the articles of *People's Daily* for the first 10 days in January 1998.

We divide the Chinese treebank into 6 sub-corpora with equivalent sizes (each about 50,000 words). The English treebank is divided into 23 sub-corpora of such sizes, where 6 are randomly chosen. This way, the influences from different sizes would be minimized. Then punctuation marks are excluded from discussion so that we can focus on word-word dependency. This is done by assigning the roles of the marks to their head words. As the sub-corpora in the same language are found to be basically homogeneous, we deem the differences between the two sets of sub-corpora as resulting from languages *per se*. E1 to E6 in English and C1 to C6 in Chinese, each boasting a size of 43,200+ words, are the resulting 12 sub-corpora. Annotated with quite similar dependency schemes, these sub-corpora are of the same size and of the same genre, and are thus eligible materials for cross-language comparison.

Besides the terms defined in Section 1, another two terms carrying great significance to dependency trees are valency and dependency distance. In dependency trees, governed by the same head/governor, sibling nodes or dependents of the same head are the "valency" of the very head. Different from the traditional meaning of valencies which cover only obligatory arguments and usually refer to verb valencies (e.g. in Köhler 2005; Gao et al. 2014), valencies here, put forward by Liu (e.g. 2009), are actually generalized valencies which cover not only obligatory arguments but also optional adjuncts of heads or valency carriers of all potential types of parts of speech. Zhang and Liu (2017) examined the distribution patterns of generalized valencies and validated their linguistic status.

"Dependency distance (DD)" is another viable necessitating more elaboration. Following the definition in Liu (2007), it is the distance between the dependent and its head in the linear sentence. For instance, the DD between "Vinken" and "old" in Fig. 1 is $|2–5|=3$, and that between "Vinken" and "Pierre" is $|2–1|=1$. Liu (2008) proposed that DD can be a metric of language comprehension difficulty. The longer the distance is, the more difficult it is to comprehend a sentence. When DD increases to a certain extent beyond the capacity of working memory, a sentence gets increasingly difficult or even beyond comprehension (Gibson 1998; Gibson & Pearlmutter 1998; Hiranuma 1999; Liu 2008).

Minimization of DD is found to be a universal feature of human language which is confined by cognitive factors (Ferrer-i-Cancho 2004, 2006; Liu 2007, 2008; Jiang & Liu 2015; Liang & Liu 2016; Wang & Liu 2017; Liu et al. 2017). Using 20 corpora of different languages, Liu (2008) formulated and testified hypotheses concerning DD, validating that the human language parser would prefer a minimized DD, which bears a threshold of 4. A combined force of grammar and cognition keeps DD within such a threshold. A similar study on 37 languages (Futrell et al. 2015) further validated the findings of Liu (2008). Research findings in computational linguistics also indicate minimization of DD can help enhance the accuracy rate of syntactic parsing.

While fitting mathematical models or functions to the data, we employ NLREG 6.3 and Altmann Fitter 3.1 for calculating parameter values and determination coefficients (R^2). In all cases, when a function is used, x and y stand for the independent and dependent variables, respectively. When we examine the interrelationship between sentence lengths and other variables, those lengths with occurrences less than 5 are eliminated to avoid a data scarcity problem.

3 Results and discussions

In this part, sequentially, each sub-section will address one or two research questions. With these questions answered, we will gradually incorporate more elements or interrelations into the suggested synergetic model. How sentence lengths influence the tree widths and heights will be first explored, followed by an examination of how dependency tree widths and heights cooperate and compete with each other. Finally, the interrelations among the widest layers, their numbers of nodes and the relevant sentence lengths will be discussed, with an aim of further expanding the synergetic model.

Before getting down to each research question, we present the types and tokens of widths and heights in both languages in Tab. 1, the data summary. The total number of tokens corresponds to that of sentences in each sub-corpus. Evidence of homogeneity among the sub-corpora of the same language abounds. Tab. 1 is a typical testimony, where for instance, average heights of the sub-corpora are quite homogeneous in one language and different from those in another language. Therefore later on in this paper, we can use E1 and C1 as representative examples for each language.

Tab. 1: A data summary (Aver. = average)

	Width types	Height types	Tokens	Aver. sentence lengths	Aver. widths	Aver. heights	Aver. DD
C1	23	14	2172	19.92	5.77	6.17	3.08
C2	21	14	2154	20.10	5.87	6.09	3.32
C3	20	12	2153	20.09	5.85	6.10	3.08
C4	21	12	2163	20.01	5.90	6.09	3.07
C5	21	13	2138	20.23	5.86	6.20	3.07
C6	20	14	2105	20.55	5.95	6.20	3.07
Aver.	21	13.20	2147.50	20.14	5.87	6.14	3.12
E1	19	20	2001	21.62	5.46	7.50	2.35
E2	18	17	1965	22.02	5.48	7.61	2.37
E3	17	17	2024	21.38	5.36	7.51	2.33
E4	18	17	2072	20.88	5.43	7.24	2.35
E5	21	21	2007	21.55	5.55	7.31	2.57
E6	15	17	2068	20.92	5.34	7.34	2.31
Aver.	18	18.20	2022.80	21.40	5.44	7.42	2.38

3.1 Sentence lengths and dependency tree widths/heights

We start from the investigation into the interrelationship between sentence lengths and dependency tree widths/heights (**Research Questions 1 & 2**).

Take C1 as an example. For simplicity, we will check their correlations first. The Pearson correlation between lengths and widths of C1 is 0.854 (p-value = 0.000) and that between lengths and heights of C1, 0.776 (p-value = 0.000). The latter result accords with Liu and Jing's (2016) finding. Similarly, the Pearson correlation between lengths and widths, and that between lengths and heights in E1 are both 0.76 (p-value = 0.000). A strong link between lengths and widths/heights in both languages is corroborated.

We would employ the average of quantities to examine the general trend, which is a common practice in quantitative linguistics. It is shown in the Appendix and its graphic representation (Fig. 3) that with sentence lengths growing, average dependency tree widths and heights grow as well.

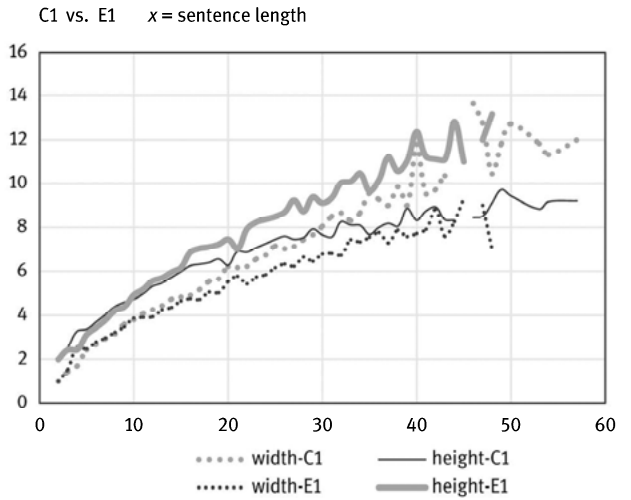


Fig. 3: Representation of the Appendix (Sentences with length frequencies less than 5 are eliminated)

Examining the correlations is not the ultimate aim as dependencies/interrelations among linguistic variables are not supposed to be linear (Köhler 2012). Typical dependencies/interrelations between two or more linguistic quantities usually take the form of a power law. We thereby fit to the data here a very well-established power law function

$$y = ax^b \tag{1}$$

where x stands for the sentence lengths, and y, tree widths/heights.

Adding a node to a tree, wherever it occurs, will simultaneously influence the generalized valency of its governor and change the dependency distance of the whole sentence.

As is shown in Tab. 1 (the summary of width and height data), with shorter mean sentence lengths, the Chinese data bear a bigger mean dependency distance. This result accords with Liu's (2008) finding. This is partly because the Chinese trees are wider but lower than English trees (Tab.1), suggesting that the growth of widths possibly exerts a bigger influence on dependency distance than the growth of heights does.

When a newly added node increases the tree height, it won't function as a sibling node of any existing node at all; in other words, it indicates a change of generalized valency of its valency carrier from 0 to 1. Simultaneously, as this

node is more likely to be close to its head in the linear sentence (than when a node increases the tree width), it will exert a lesser influence on dependency distance.

In contrast, when the new node increases the tree width, it is much more likely for its governor to have an already existent generalized valency and hence it is much more likely for it to be a non-neighboring word of the governor in the linear sentence, thus generating a greater dependency distance.

Semantic, pragmatic and syntactic necessities increase generalized valency of a word to a certain degree; conversely, they also restrain such growth. Likewise, as is mentioned before, human memory capacity both allows a certain distance between dependents and their heads and stops it from getting too long to be comprehensible (Gibson 1998; Gibson & Pearlmutter 1998; Hiranuma 1999; Liu 2008).

So in Function (1), we hypothesize that Parameter a is related to the initial value of the sequence (but not necessarily 1 or 2 in the real data), and b, a combinational force of the resistance from the growth of the generalized valencies and that from the growth of dependency distances. When y in Function (1) stands for heights, the resistance from increasing generalized valencies seems to be more at play; when y stands for widths, the resistance from increasing dependency distance seems to assume more weight.

For the two afore-mentioned types of resistance, we illustrate with C1 and E1. The resistance from the growth of generalized valencies can be displayed by their distributions (Tab. 2). Obviously, nearly half of words go without any dependents; a little less than 1/3 bear one dependent; and the percentage of words with two dependents drops dramatically to 10.57 and 15.10 in C1 and E1, respectively. This general trend persists when valency grows, indicating increasing difficulty to bear more dependents.

Likewise, the resistance from the growth of dependency distances can be exhibited through their distributions (Tab. 3). To save space, we only display the top ten, which can clearly present the trend of mounting pressure from increasing dependency distance. This tendency is in agreement with Liu's (2007) finding.

These two trends in the resistance can also be more readily visible in Fig. 4, which visually presents the data of Tab. 2 and Tab. 3. It's interesting to observe that the initial drop of dependency distance is faster than that of valencies. Later in the sequence, they seem to drop more slowly at a similar rate.

Tab. 2: Generalized valency distributions of C1 and E1

C1			E1		
Valency	**Frequency**	**%**	**Valency**	**Frequency**	**%**
0	20054	46.36	0	19571	45.24
1	13847	32.01	1	12730	29.42
2	4573	10.57	2	6533	15.10
3	2503	5.79	3	2865	6.62
4	1346	3.11	4	1110	2.57
5	616	1.42	5	342	0.79
6	223	0.52	6	80	0.18
7	66	0.15	7	25	0.06
8	17	0.04	8	6	0.01
9	8	0.02	9	2	0.00
10	3	0.01			
11	1	0.00			
29	1	0.00			
39	1	0.00			
Total	43259			43264	

Tab. 3: Generalized valency distributions of C1 and E1 (top 10)

C1			E1		
DD	**Frequency**	**%**	**DD**	**Frequency**	**%**
1	20814	50.66	1	22289	54.02
2	6870	16.72	2	8538	20.69
3	3715	9.04	3	4177	10.12
4	2356	5.74	4	2081	5.04
5	1615	3.93	5	1116	2.71
6	1132	2.76	6	719	1.74
7	898	2.19	7	460	1.12
8	672	1.64	8	393	0.95
9	521	1.27	9	287	0.70
10	402	0.98	10	205	0.50

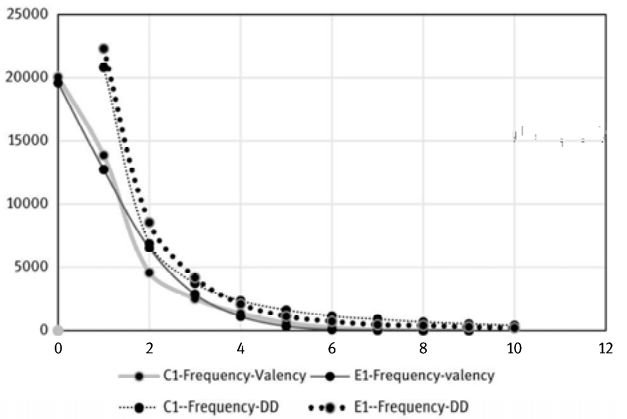


Fig. 4: Graphic representation of Tab. 2 and Tab. 3

Fitting Function (1) to the data yields excellent results as presented in Tab. 4 and its graphic representation Fig. 5. The values of Parameter *a* are different between widths and heights as dependent variables, as the initial value of widths is 1 and that of heights, 2. All values of the same parameters (including *a* and *b*) are grouped together in the same language but separated from another language, indicating similarities within one language and differences between languages. The *b* values in the Chinese data differ when the dependent variable changes, indicating the resistances carry different weight on heights and on widths for the language. This can be a tentative reason for the general differences exhibited in Tab. 1.

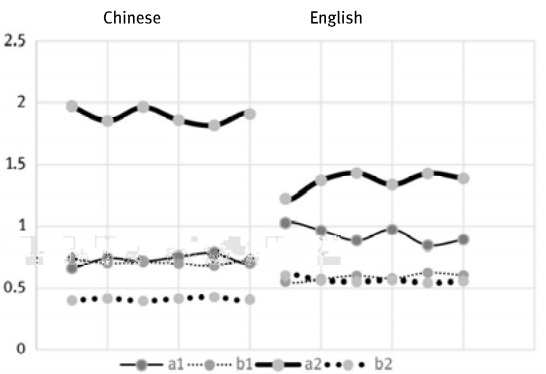


Fig. 5: Graphic representation of Tab. 4

Tab. 4: Fitting Function (1) to the widths and heights data (complete data)

	Width as dependent variable			Height as dependent variable			
	R^2	a_1	b_1	R^2	a_2	b_2	b_1+b_2
C1	0.9640	0.66	0.74	0.9716	1.97	0.40	1.14
C2	0.9838	0.74	0.70	0.9694	1.85	0.41	1.11
C3	0.9836	0.72	0.71	0.9598	1.96	0.39	1.10
C4	0.9839	0.76	0.70	0.9809	1.86	0.41	1.11
C5	0.9802	0.79	0.68	0.9766	1.81	0.42	1.10
C6	0.9761	0.70	0.72	0.9551	1.91	0.40	1.12
E1	0.9669	1.02	0.55	0.9828	1.22	0.60	1.15
E2	0.9863	0.96	0.57	0.9855	1.37	0.56	1.13
E3	0.9824	0.88	0.60	0.9780	1.43	0.55	1.15
E4	0.9722	0.97	0.57	0.9732	1.33	0.57	1.14
E5	0.9617	0.84	0.62	0.9686	1.43	0.54	1.16
E6	0.9800	0.89	0.60	0.9763	1.39	0.56	1.16

With the interrelations established, we update Fig. 2 and come up with Fig. 6.

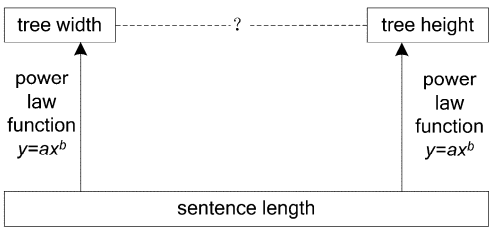


Fig. 6: Establishing the link between sentence lengths and tree widths/heights

For a circuit to be complete, we need to resume with the interrelation between tree widths and tree heights (**Research Question 3**), which will be tackled in Section 3.2.

3.2 Dependency tree widths vs. heights

Fig. 7(a) and 7(b) exemplify the height-width interrelations in E1. Fitting the power law function (1) yields determination coefficients of around 0.77. These

fitting results are still acceptable, but insufficient enough to establish a strong and convincing link.

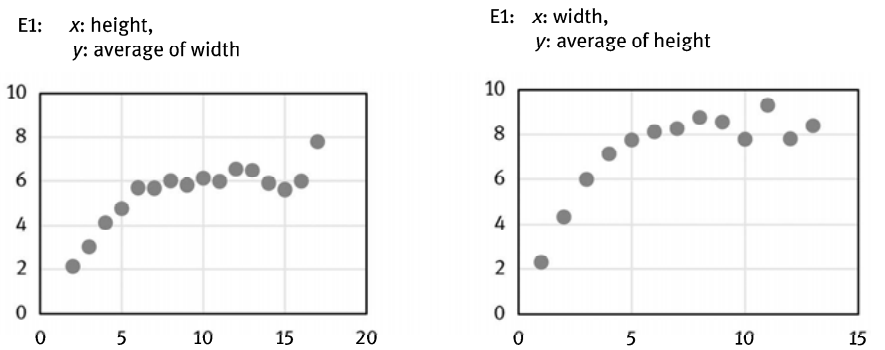


Fig. 7: Widths and heights, **a.** heights and average of widths, **b.** widths and average of heights

The curves in Fig. 7 are different from the power law or Zipf's law-related distributions, but very identical to Piotrowski Law, which is generally a developmental law. The Piotrowski Law, in the form of a logistic model, is a curve well-known for modeling various linguistic phenomena (e.g. Altmann 1983; Leopold 2005). This function typically models linguistic changes over time. We posit that the function applies to the growth of dependency tree widths/heights as well.

We thus fit to the data the logistic function

$$y = c/(1 + e^{-bx}) \tag{2}$$

As presented in Tab. 5, the fittings are good enough for us to update Fig. 6 and come up with Fig. 8, where a complete circuit of relations among variables is seen. We claim this is a tentative synergetic model between linear sentences and two-dimensional trees. The double-headed arrows indicate a mutual influence between widths and heights.

We postulate that a minimization of dependency distance (cf. Section 2) is achieved through the competition and cooperation between tree width growth (thereby resisting the growth of tree heights) and tree height growth (thereby resisting the growth of tree widths). An interesting observation in Tab. 4 is that for the same sub-corpus, all the values of b_1+b_2 in Chinese approximate 1.11, and those in English, 1.15. As b stands for a combinational force resisting the growth of the generalized valencies and of dependency distances, the sum being a con-

stant indicates that the sum of the two opposing forces is constant. This interesting finding suggests the strong interrelations (a trade-off resulting from both competition and cooperation) between the two forces. This, we propose, constitutes an important cause for minimization of dependency distance of the whole sentence and for the balance of dependency trees.

Tab. 5: Fitting the Piotrowski Law to data of widths and heights (complete data)

	Width as dependent variable					Height as dependent variable			
	R^2	*a*	*b*	*c*		R^2	*a*	*b*	*c*
C1	0.9533	2.25	0.36	8.33	C1	0.9780	11.10	0.57	8.20
C2	0.9742	2.12	0.30	8.64	C2	0.9852	9.55	0.45	9.73
C3	0.9742	2.36	0.38	8.07	C3	0.9758	10.18	0.54	8.54
C4	0.9872	1.97	0.29	8.51	C4	0.9199	9.43	0.49	9.20
C5	0.9639	2.07	0.30	8.73	C5	0.9903	7.61	0.37	10.41
C6	0.9591	2.35	0.41	7.98	C6	0.9948	8.13	0.50	8.62
E1	0.9615	6.00	0.89	8.42	E1	0.8656	6.63	0.62	6.31
E2	0.9653	6.93	0.95	8.52	E2	0.9478	9.07	0.75	5.97
E3	0.9915	5.52	0.95	8.28	E3	0.8912	11.09	0.80	5.81
E4	0.9743	5.72	0.93	8.08	E4	0.9155	11.30	0.87	5.83
E5	0.8031	5.85	1.02	7.91	E5	0.9064	11.05	0.82	6.04
E6	0.9766	7.84	1.04	8.04	E6	0.8433	12.99	0.88	5.79

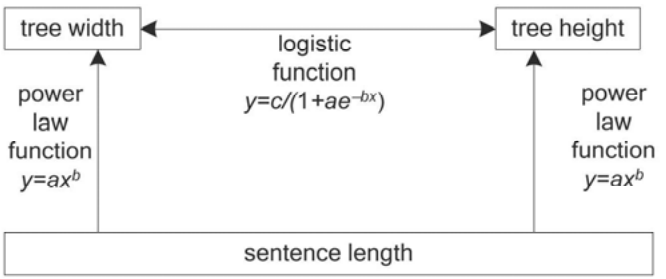


Fig. 8: Establishing the link between dependency tree widths and heights

Another interesting observation is that for sentences with less than 16 words, the widths (dotted lines) in both languages grow at a similar rate (Fig. 3). When sentences grow longer, Chinese dependency trees begin to grow wider at a faster rate. Both languages share a similar height (solid lines) for the same sentence length when it's shorter than 16 words. When sentences grow longer, contrary to the trend of width growth, heights in English grow generally faster than those in Chinese. This also contributes to the general trends exhibited in Tab. 1. The cause of this contrast has much to do with the generalized valencies of different parts of speech in the two languages, which will be addressed in a parallel study of this project.

We won't end up here with Fig. 8. Rather, in the next sub-section, we will address the fourth research question and examine whether the widest layer can also be incorporated into the model.

3.3 Widest layers vs. their widths

In the example (Fig. 1),

$$\begin{gathered} 3(\text{the widest layer}) \\ \times 5(\text{the width of the widest layer}) \\ \hline = 15(\text{the sentence length}) \end{gathered} \tag{2}$$

Now we examine whether the connection (Sentence length = tree width * widest layer) persists in the sub-corpora (**Research Question 4**).

The Pearson correlation of sentence lengths and the product of tree width and widest layer is found to be 0.859 in E1, suggesting a direct link. We thus fit Function (3) to all the data

$$y = ax - b \tag{3}$$

where x is the sentence length, and y, the product of average widest layers and average widths of such sentence lengths. We use such a linear function for fitting as there are actually three variables rather than two since the product is a result of two variables. The interrelations between any two are still non-linear.

The fitting results are presented in Tab. 6. Fig. 9 presents the fitting of E1 data. With all R^2 going above 94.4%, we deem the fittings excellent.

The cause for such regularity is still unknown and it's worth more investigation in future research endeavors. We propose that it results from the competition and cooperation of the growth of widths and heights, or in other words, the

competition between changes of dependency distances and of generalized valencies. It is at the widest layer that the widths and heights strike a balance between the two opposing forces.

Tab. 6: Fitting the linear function to the data

	a	*b*	R^2		*a*	*b*	R^2
C1	1.065	1.3420	0.979	E1	1.105	1.158	0.981
C2	1.004	0.7145	0.982	E2	1.106	1.716	0.977
C3	1.029	1.0030	0.972	E3	1.132	1.935	0.977
C4	1.062	1.4750	0.990	E4	1.100	1.359	0.944
C5	1.008	0.4925	0.973	E5	1.140	2.263	0.949
C6	1.038	0.8509	0.986	E6	1.066	1.051	0.975

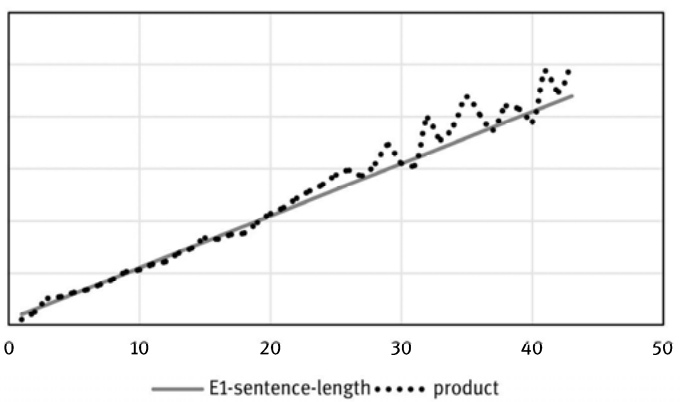


Fig. 9: Fitting the linear function to E1 data

The change of dependency distances is, in turn, influenced by the cognitive constraints of the brain. The valencies are a result of two forces at play: the semantic, grammatical and pragmatic requirements and the restraint of dependency distance. All these factors work together to contribute to the balanced shape of dependency trees.

Based on the research findings in this part, a final synergetic model is built (Fig. 10).

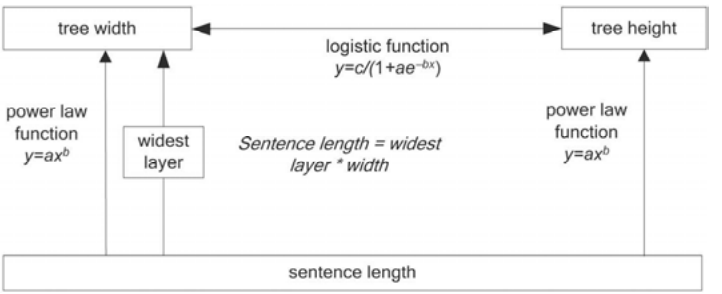


Fig. 10: Incorporating the widest layer into the model

4 Concluding remarks

Choosing two similarly annotated dependency treebanks of the news genre, one in Chinese and one in English, this paper divides the treebanks into 6 sub-corpora of equivalent sizes in each language, and examines the interrelations between the linear sentences and the two-dimensional dependency trees. Basically, each set of sub-corpora behaves homogeneously in the very language.

This study yields the following research findings.

a) When sentences grow longer, the dependency trees grow higher, generating more layers. A power law function can excellently capture the interrelation between tree heights and sentence lengths.
b) Similarly, with the growth of sentence lengths, the dependency trees grow wider, including more dependents at a certain layer. The previous power law function can also reflect the interrelation between tree widths and sentence lengths.
c) A logistic function, which is also employed as the Piotrowski law, captures the link between dependency tree widths and heights.
d) The layer which has the most nodes grows with sentence lengths as well. The product of the widest layer and its width approximates the sentence length.
e) Based on the previous research findings, a tentative syntactic synergetic model is built.

This study is, to our knowledge, the first empirical attempt at establishing a synergetic syntactic model based on the framework of dependency grammar. Future research endeavours may further corroborate the findings of this paper and expand the model to incorporate more properties (over 20) with a hope of

establishing a more comprehensive synergetic model. For instance, since both valencies and dependency distances play a role in the relationship among the variables (Sections 3.1 and 3.2), a question spontaneously arises: can we also incorporate these two variables into the synergetic model? This question will be addressed in further studies.

What needs to be pointed out, however, is that using merely two languages and one genre means that the model is not universal enough. For more universality, further research efforts might cover more dependency annotation schemes (e.g. universal dependency) and include a wider coverage of language varieties, both in terms of language typologies and genres.

Acknowledgement: This work was supported by the Fundamental Research Funds for the Central Universities (of Zhejiang University) under Grant No. 2016XZA103. Sincere thanks also go to Haiqi Wu, who helped with data for the research.

References

Altmann, Gabriel. 1983. Das Piotrowski-Gesetz und seine Verallgemeinerungen. In Karl-Heinz Best & Jörg Kohlhase (Eds.), *Exakte Sprachwandelforschung* (pp. 54–90). Göttingen: Herodot.

Ferrer-i-Cancho, Ramon. 2004. Euclidean Distance between Syntactically Linked Words. *Physical Review E, 70* (5), 056135.

Ferrer-i-Cancho, Ramon. 2006. Why Do Syntactic Links Not Cross? *ELP (Europhysics Letters), 76* (6), 1228–1234.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. *PNAS, 112* (33), 10336–10341.

Gao, Song, Hongxin Zhang & Haitao Liu. 2014. Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics, 21*(1), 1–21.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1), 1–76.

Gibson, Edward & Neal J. Pearlmutter. 1998. Constraints on sentence comprehension. *Trends in Cognitive Sciences, 2* (7), 262–268.

Hajič, Jan, Eva Hajičová, Jarmila Panevová, Petr Sgall, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský. 2012. Prague Czech-English Dependency Treebank 2.0 LDC2012T08. DVD. Philadelphia: Linguistic Data Consortium.

Haken, Hermann. 1978. *Synergetics*. Heidelberg: Springer.

Hiranuma, So. 1999. Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics*, 11, 309–322.

Jiang, Jingyang & 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.

Jing, Yingqi & Haitao Liu. 2015. Mean Hierarchical Distance Augmenting Mean Dependency Distance. In Proceedings of Third International Conference on Dependency Linguistics (Depling 2015) (pp. 161–170). Uppsala University, Uppsala, Sweden. August 24–26.

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, Reinhard. 2005. Quantitative untersuchungen zur valenz deutscher verbena. *Glottometrics*, 9, 13–20.

Köhler, Reinhard. 2012. *Quantitative Syntax Analysis*. Berlin: de Gruyter Mouton.

Köhler, Reinhard & Gabriel Altmann. 2000. Probability distributions of syntactic units and properties. *Journal of Quantitative Linguistics, 7* (1), 189–200.

Köhler, Reinhard & Gabriel Altmann. 2009. *Problems in Quantitative Linguistics* (Vol. 2). Lüdenscheid: RAM-verlag.

Leopold, Edda. 2005. Das Piotrowski-Gesetz. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp. 627–633). Berlin/New York: Walter de Gruyter.

Liang, Junying & Haitao Liu. 2016. Interdisciplinary studies of linguistics: Language universals, human cognition and big-data analysis. *Journal of Zhejiang University (Humanities and Social Sciences)*, 1, 108–118.

Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics*, 15, 1–12.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.

Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.

Liu, Haitao. 2017. Distribution of hierarchical sentence structures. *Foreign Language Teaching and Research, 49*(3), 345–352.

Liu, Haitao & Yingqi Jing. 2016. A quantitative analysis of English hierarchical structure. *Journal of Foreign Languages, 39*(6), 2–11.

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

Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz & Ann Taylor. 1999. The Penn Treebank 3. Linguistic Data Consortium, Catalog #LDC99T42.

Qiu,Likun,Yue Zhang, Peng Jin & Houfeng Wang. 2014. Multi-view Chinese treebanking. In Proceedings of the 25th International Conference on Computational Linguistics (COLING) (pp. 257–268). Dublin, Ireland, August.

Tesnière, Lucien. 1959. *Éléments de Syntaxe Structural*. Paris: Klincksieck.

Wang, Hua & Haitao Liu. 2014. The effects of length and complexity on constituent ordering in written English. *Poznań Studies in Contemporary Linguistics, 50*(4), 477–494.

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

Zhang, Hongxin & Haitao Liu. 2017. Motifs of Generalized Valencies. In Liu, Haitao & Junying Liang (Eds.), *Motifs in Language and Text* (pp. 231–260). Berlin: de Gruyter Mouton.

Ziegler, Arne & Gabriel Altmann. 2003. Text Stratification. *Journal of Quantitative Linguistics, 10*(3), 275–292.

Appendix

Sentence lengths, their average dependency tree widths and heights (C1 and E1)

C1				E1			
Length	Frequency	Average width	Average height	Length	Frequency	Average width	Average height
2	35	1.00	2.00	2	14	1.00	2.00
3	42	1.45	2.55	3	14	1.57	2.43
4	64	1.72	3.28	4	25	2.56	2.44
5	69	2.43	3.36	5	25	2.48	3.12
6	58	2.69	3.69	6	32	2.78	3.44
7	66	2.94	4.03	7	39	2.97	3.79
8	62	3.16	4.39	8	47	3.23	4.21
9	60	3.60	4.58	9	34	3.47	4.38
10	84	3.79	4.71	10	45	3.87	4.91
11	69	4.09	4.97	11	52	3.90	5.17
12	78	4.24	5.31	12	69	3.93	5.55
13	85	4.42	5.46	13	67	4.24	5.69
14	84	4.73	5.71	14	79	4.35	5.96
15	67	4.82	5.99	15	77	4.64	6.17
16	58	4.90	6.26	16	73	4.74	6.88
17	74	5.16	6.32	17	65	4.71	7.06
18	69	5.54	6.41	18	75	5.04	7.12
19	75	5.64	6.56	19	83	5.01	7.24
20	65	6.22	6.28	20	90	5.52	7.47
21	71	6.17	6.92	21	62	5.76	7.03
22	46	6.22	6.89	22	76	5.42	7.93
23	77	6.57	7.09	23	86	5.73	8.24
24	55	6.75	7.27	24	73	5.84	8.37
25	54	7.17	7.46	25	70	6.17	8.49
26	48	7.04	7.60	26	60	6.33	8.70
27	47	7.15	7.47	27	56	6.23	9.23
28	51	7.45	7.55	28	58	6.66	8.72
29	40	7.68	7.93	29	50	6.46	9.38
30	36	8.06	7.67	30	40	6.80	9.10
31	40	8.50	7.58	31	38	6.82	9.34

C1				E1			
Length	Frequency	Average width	Average height	Length	Frequency	Average width	Average height
32	26	8.65	8.27	32	35	6.74	10.03
33	26	8.31	8.12	33	33	7.45	10.09
34	21	8.67	8.10	34	26	7.35	10.46
35	29	9.55	7.69	35	28	7.57	9.64
36	27	9.26	8.00	36	31	7.77	10.16
37	22	9.00	8.18	37	23	7.30	11.22
38	19	9.84	8.05	38	16	7.88	10.56
39	21	9.00	8.86	39	22	7.59	11.05
40	15	11.80	8.33	40	11	7.73	12.36
41	8	9.50	8.75	41	13	7.92	11.23
42	14	9.71	8.93	42	17	8.82	11.12
43	14	10.36	8.36	43	10	7.60	11.10
44	9	10.44	8.33	44	11	8.27	12.82
45	3	10.67	8.67	45	6	9.33	11.00
46	11	13.64	8.45	46	2	6.00	13.50
47	8	12.75	8.50	47	5	9.00	12.00
48	9	10.44	9.11	48	6	7.00	13.17
49	7	11.86	9.71	49	4	7.75	13.50
50	7	12.71	9.43	50	2	9.50	10.00
51	1	13.00	9.00	51	4	12.00	10.75
52	2	10.00	10.00	52	1	8.00	13.00
53	6	11.83	8.83	53	1	11.00	10.00
54	6	11.33	9.17	54	3	11.67	12.67
55	2	12.50	9.50	55	3	11.00	10.67
56	4	11.75	9.75	56	2	12.50	9.50
57	5	12.00	9.20	57	1	14.00	6.00
58	2	11.50	9.00	58	1	9.00	13.00
59	1	18.00	8.00	59	1	11.00	15.00
60	3	12.00	10.33	60	2	11.50	12.50
61	1	18.00	7.00	62	1	8.00	23.00
62	1	11.00	14.00	63	1	15.00	11.00
63	2	13.50	11.00	73	2	12.00	14.50
66	2	15.00	9.50	80	1	25.00	6.00
67	1	15.00	10.00	89	1	29.00	11.00

C1				E1			
Length	**Frequency**	**Average width**	**Average height**	**Length**	**Frequency**	**Average width**	**Average height**
69	1	14.00	12.00	92	1	18.00	13.00
74	2	13.50	11.50				
77	1	15.00	10.00				
78	1	15.00	10.00				
85	1	24.00	8.00				
89	1	22.00	8.00				
106	1	14.00	17.00				
Sum	2172			**Sum**	2001		
Aver.		5.77	6.17	**Aver.**		5.46	7.50

Radek Čech*, Jiří Milička, Ján Mačutek, Michaela Koščová, Markéta Lopatková

Quantitative Analysis of Syntactic Dependency in Czech

Abstract: The article presents a quantitative analysis of some syntactic dependency properties in Czech. A dependency frame is introduced as a linguistic unit and its characteristics are investigated. In particular, a ranked frequencies of dependency frames are observed and modelled and a relationship between particular syntactic functions and the number of dependency frames is examined. For the analysis, the Czech Universal Dependency Treebank is used.

Keywords: syntactic dependency; dependency frame; syntactic function

1 Introduction

A hierarchical structure of a sentence can be expressed by the dependency grammar formalism (Mel'čuk 1998; Hudson 2007). This formalism describes the structure of a sentence in a form of a tree graph: nodes of the graph represent words, while links between nodes represent syntactic relationships between words. Within the approach, there is a syntactic function assigned to each word in a sentence, e.g., predicate, subject, object, determiner, etc. The dependency formalism allows us to gain an insight into relationships among sentence elements and it represents widely accepted formalism for description of syntactic properties of language. In this study, we have decided to adopt this approach for an analysis of more general properties of syntax; namely, we focus on the frequency distribution of so called dependency frames (for details, see Section 2) as well as on the relationship between the frequency of syntactic functions (subject, predicate, etc.) and the number of dependency frames of particular units. This kind of analysis is based on an assumption that a regular frequency

Radek Čech, University of Ostrava, Ostrava, Czech Republic, cechradek@gmail.com
Jiří Milička, Charles University, Prague, Czech Republic
Ján Mačutek, Comenius University in Bratislava, Bratislava, Slovakia
Michaela Koščová, Comenius University in Bratislava, Bratislava, Slovakia
Markéta Lopatková, Charles University, Prague, Czech Republic

https://doi.org/10.1515/9783110573565-003

distribution of language units can be interpreted as a result of very general facets of human language behavior. Specifically, the regular distribution is explained as a consequence of the least effort principle (Zipf 1949) or as an outcome of a diversification process in the language (Altmann 2005). Moreover, the regular distribution as well as the relationship between the frequencies of given units play a fundamental role in the synergetic model of language, which makes it possible to describe and, most importantly, to explain mutual interrelations among various language properties (Köhler 1986, 2005, 2012). Based on an expectation that relationships between syntactic functions are ruled by the same general mechanisms as other language properties, we set up the following hypotheses:

1) there is a regular frequency distribution of dependency frames in general in a language;
2) there is a regular frequency distribution of dependency frames for each syntactic function; differences among the distributions of individual syntactic functions are caused by their specific syntactic properties; differences are manifested by different models or different parameter values in the same model;
3) the more frequent the syntactic function, the more dependency frames it has.

This study represents a further step in the endeavor to apply quantitative methods in the dependency syntax analysis (e.g., Liu et al. 2017; Liu 2009; Mačutek et al. 2017; Čech et al. 2017).

The article is organized as follows. Section 2 introduces main characteristics of dependency frame. Section 3 describes the language material and the methodology. Section 4 discusses the results of the study and Section 5 concludes the article.

2 Dependency frame

According to the dependency grammar formalism, the (surface) syntactic structure of sentence (1) can be expressed by the tree graph displayed in Fig. 1.

(1) Christiane gave you a good answer

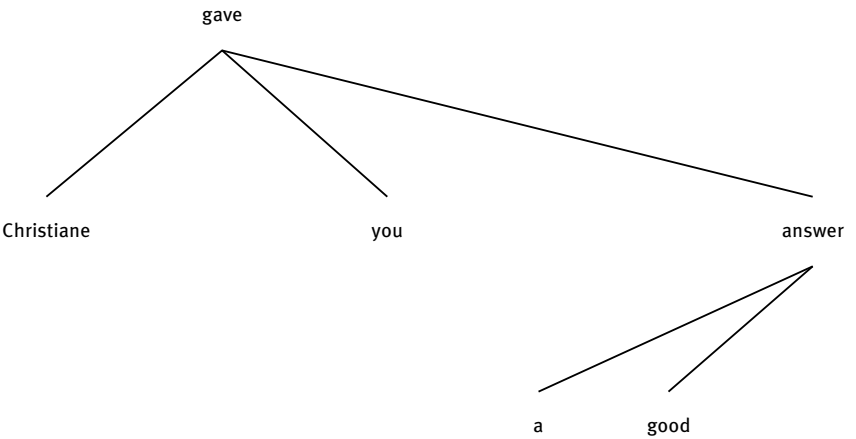


Fig. 1: The (surface) syntactic structure of sentence (1).

Following the approach presented by Čech et al. (2010), we set up a dependency frame (hereafter DF) as a basic unit of our analysis. Specifically, the DF is defined as a set of syntactic functions assigned to all words directly dependent on a given node in the tree. Particular syntactic functions are determined in accordance with the annotation used in the Universal Dependencies project (UDs)[1] (Zeman 2015; Nivre et al. 2016). The syntactic annotation within the UDs is based on the so called Stanford dependencies (de Marneffe et al. 2006, 2008, 2014). For illustration, the UDs annotation of sentence (1) is presented in Fig. 2. Syntactic functions (within the UDs project, they are called syntactic relations) are assigned to links between pair of words; they display the syntactic function of a word to which an arrow points. Thus, *Christiane* represents the nominal subject (nsubj), *you* represents the indirect object (iobj), *a* represents the determiner (det), etc.

There are two frame evoking words in sentence (1), *gave* and *answer,* and thus two DFs are identified there. In particular, the predicate *gave* has the frame:

[nsubj; iobj; obj]

1 http://universaldependncies.org/

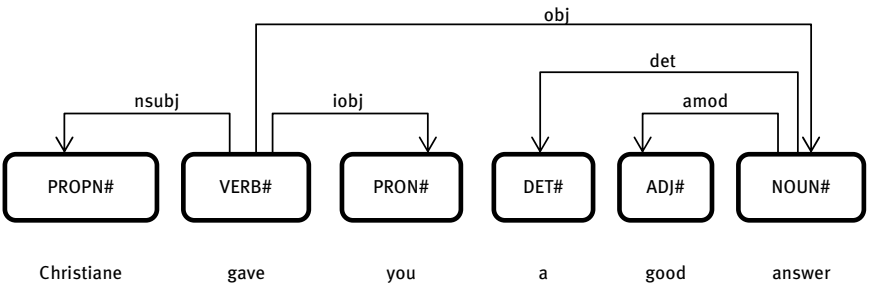


Fig. 2: The annotation of sentence (1) based on the principles used in the UDs project.

as the words *Christiane*, *you* and *answer* are directly dependent on the word *gave* (punctuation is disregarded here, see below); similarly, the object *answer* has the frame:

[det; amod]

as the words *a* and *good* are annotated as directly dependent on the word *answer* according to the UDs principles. All other words in sentence (1) have no directly dependent words, thus, no other DF can be determined there. In general, the notion of the frame evoking word (hereafter FEW) denotes a word which is a parent of DF elements in a syntactic tree; in other words, FEWs are all non-terminal elements of syntactic tree.

It should be noticed that word order is not taken into account as we are focusing on Czech, the language with high degree of word order freedom, where syntactic functions are represented by other means (esp. morphology) than the word order.

3 Language material and methodology

The Czech Universal Dependency Treebank, based on the Prague Dependency Treebank 3.0 (Bejček et al. 2013), is used in this study. The treebank consists of 87,913 sentences and about 1.5 million words/tokens. Its domain is mainly newswire, reaching also to business and popular scientific articles from the 1990s.[2]

2 see https://github.com/UniversalDependencies/UD_Czech/blob/dev/README.md

The procedure briefly sketched in Section 2 was applied to all sentences in the corpus. We used syntactic functions for determining DFs presented in Tab. 1.

In the analysis, we employed 21 syntactic functions (out of the total of 37 functions used in the UDs annotation scheme). We have omitted technical functions (such as punctuation, unspecified dependency, etc.) and linguistically not well-established functions (such as flat multiword expression, orphan, etc.). In sum, we have followed the UDs approach according to which syntactic functions presented in Tab. 1 represent dependency relations in their narrow sense (cf. http://universaldependencies.org/u/dep/index.html).

Tab. 1: The list of syntactic functions used for the determination of DF.

Syntactic function (relation)	Abbreviation
nominal subject	nsubj
object	obj
indirect object	iobj
oblique nominal	obl
vocative	vocative
nominal modifier	nmod
appositional modifier	appos
numeric modifier	nummod
clausal subject	csubj
clausal complement	ccomp
open clausal complement	xcomp
adverbial clause modifier	advcl
adverbial modifier	advmod
discourse element	discourse
clausal modifier of noun (adjectival clause)	acl
adverbial modifier	amod
copula	cop
determiner	det
case marking	case
auxiliary	aux
marker	mark

The UDs annotation was also used to determine syntactic functions of FEWs. However, it seems reasonable to slightly modify the original annotation scheme for predicates since they are not explicitly annotated in the UDs. Namely, as for main clauses, we assign the predicate function (pred) to the root node of the tree represented by a verb, see sentence (2), or to the non-verb root node on which a word with the "auxiliary" (AUX) POS tag is directly dependent, see sentence (3). Thus, in sentence (2) we determine the verb *comes* as the predicate, see Figure 3,

(2) From the AP comes this story

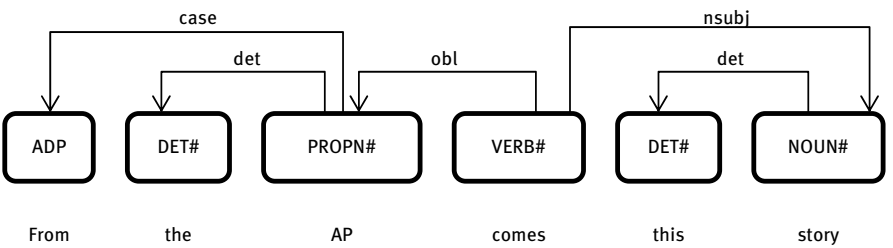


Fig. 3: The annotation of the sentence (2) based on the annotation used in the UDs project.

and in sentence (3) *young* is determined as the predicate (the auxiliary verb *was* is directly dependent on this word), see Figure 4.

(3) I was very young

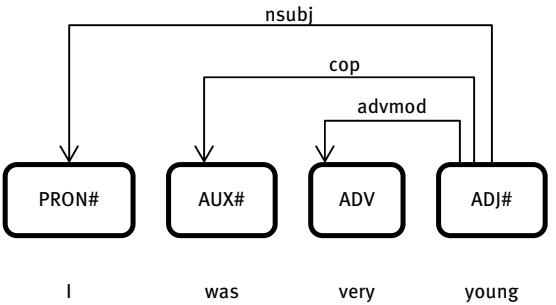


Fig. 4: The annotation of sentence (3) based on the annotation used in the UDs project.

Further, according to the UDs principles, predicates of dependent clauses are annotated in accordance with the function of the whole clause, as e.g. the clausal subject for the predicate of a subject clause, the adverbial clause modifier for the predicate of an adverbial clause (this is used also for determining DFs, see Tab. 1). In these cases, we assign the predicate function (pred) to words with syntactic functions csubj, ccomp, xcomp, advcl, and acl (which correspond to the root of clauses, in fact) if they are represented by verbs (as in sentence (4)), or to non-verb nodes on which a word with the "auxiliary" (AUX) POS tag is directly dependent. Consequently, in sentence (4) we determine the word *prove* (i.e., the verbal root of the sentence) as the predicate of the main clause, and *lost* (the verb with assigned syntactic function 'accom') as the predicate of the dependent clause, see Fig. 5.

(4) Today's incident proves that Sharon has lost his patience

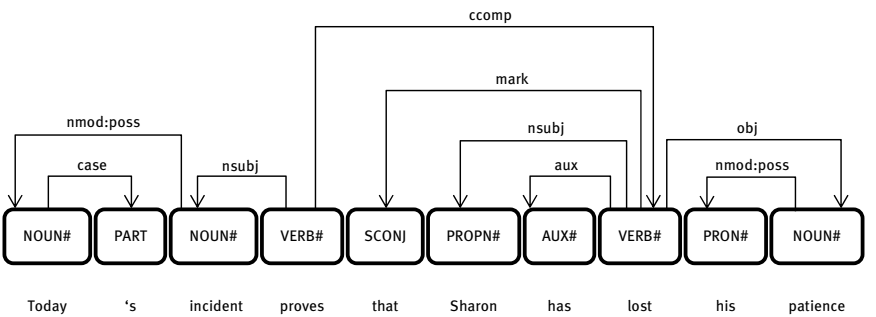


Fig. 5: The annotation of sentence (4) based on the annotation used in the UDs project.

All frequencies and numbers of occurrences were modelled by function

$$y = ax^b \tag{1}$$

There are several reasons for the choice of this particular mathematical model. First, it fits the data sufficiently well, as values of the determination coefficient are mostly greater than 0.9 (which is usually accepted as one of the criteria for a satisfactory fit in mathematical modelling of linguistic data), or at least not much lower (cf. Mačutek and Wimmer 2013 for a discussion of several goodness-of-fit measures applied in quantitative linguistics). Second, this model is very simple, nevertheless it reveals some differences among syntactic functions (see

Section 4). Finally, it is a special case of a very general mathematical formula expressing many language laws, which was presented by Wimmer and Altmann (2005). We thus remain within the general theoretical framework of quantitative linguistics, which makes it possible to investigate interrelations between syntax and other linguistic levels in future.

4 Results

The first hypothesis states that there should be a regular frequency distribution of DFs in the corpus. We have operationalized the hypothesis in two ways. First, all particular DFs and their frequencies were determined. Applying this procedure, we obtain 1,214 different DFs. The ten most frequent DFs are presented for illustration in Tab. 2 - not surprisingly, the most frequent DFs consists of one dependent word; the most frequent DFs with two dependents is represented by rank 4.[3]

Tab. 2: The ten most frequent DFs in the corpus.

Rank	Dependency frame	Frequency
1	nmod	5063
2	case	4569
3	amod	4454
4	case nmod	2262
5	amod case	2176
6	amod nmod	1613
7	obj	1117
8	det	1102
9	advmod	943
10	obl	864

Fitting function (1) to the data, we obtain parameter values a = 6375.13, b = −0.8737, with the determination coefficient R^2 = 0.897. Thus, our data corrobo-

3 All the data used for the experiments can be found at:
http://www.cechradek.cz/data/Cech_etal_Q_analysis_of_synt_dependency_results.zip

rate the first hypothesis as the value of determination coefficient is satisfactory. This result means that DF is a language unit which displays the rank-frequency distribution similar to distributions of the majority of well-established language units, such as words, lemmas, syllables, etc.

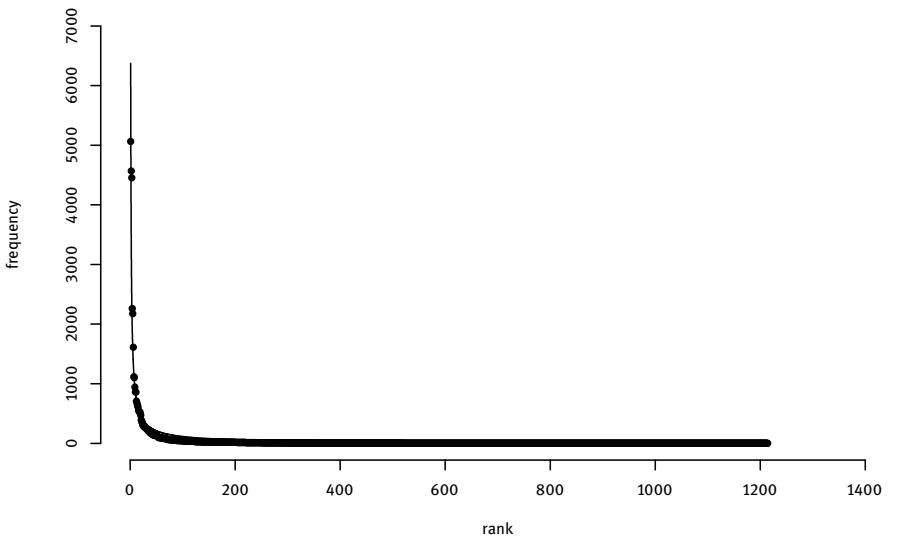


Fig. 6: Ranked frequencies of DFs in the corpus fitted with function (1).

As for the second way of the operationalization of the first hypothesis, we counted the number of all occurrences of DFs (i.e., their frequency in the data) for each chosen syntactic function as well as the number of unique DFs (i.e., the number of their types in the data). The results are presented in Tab. 3 and Fig.7 and Tab. 4 and Fig. 8, respectively.

Tab. 3: Frequency of all occurrences of DFs for each syntactic function of FEW in the data.

Rank	Syntactic function of FEW	Frequency of all DFs
1	pred	13973
2	nmod	10817
3	obl	8400
4	nsubj	5258

Rank	Syntactic function of FEW	Frequency of all DFs
5	obj	4451
6	amod	1361
7	appos	600
8	advmod	356
9	iobj	218
10	nummod	123
11	det	74
12	mark	7
13	vocative	7
14	discourse	1

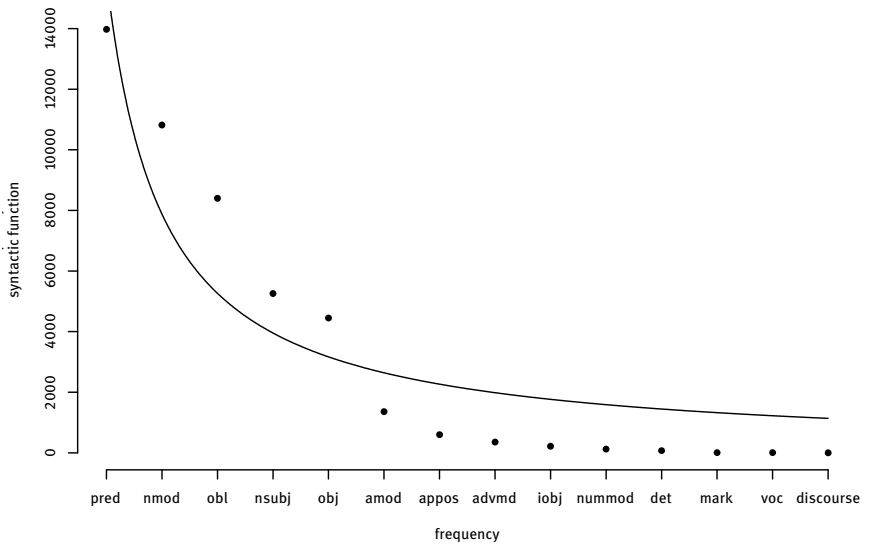


Fig. 7: Frequency of all occurrences of DFs for each syntactic function of FEW. The line represents function (1) with parameters $a = 15687.85$, $b = -0.9941$.

Fitting function (1) to the data from Tab. 3, we get following results: $a = 15687.85$, $b = -0.9941$, $R^2 = 0.849$, which is considered as acceptable result and thus the hypothesis is not rejected. However, the value of $R^2 < 0.9$ can/must be interpreted as a sign of an existence of fluctuations, which can be caused by

various reasons (e.g., the way of annotation, the character of the corpus) or/and as a sign of an unsuitability of the model used for the analysis. It is needless to say that only further research can reveal reasons for this phenomenon.

Fitting function (1) to the data from Tab. 4, we get $a = 929.42$, $b = -1.8313$, with $R^2 = 0.959$. In this case, we obtain very satisfactory fit in terms of the determination coefficient, thus, the hypothesis can be tentatively considered as corroborated, too.

The analysis of results for both frequency distributions (i.e., all DFs occurrences and number of unique DFs, i.e. types), a specific position of predicate is revealed. It can be explained as a consequence of its central role in a clause structure and its specific position in the syntactic tree - it is always the root of the tree (or the root of the subtree representing a dependent clause) and, consequently, it is not influenced by hierarchically higher syntactic elements.

Tab. 4: Number of unique DFs (DF types) for each syntactic function of FEWs.

Rank	Syntactic function of FEW	Number of unique DFs (DF types)
1	pred	947
2	nmod	146
3	obl	138
4	obj	136
5	nsubj	132
6	appos	112
7	amod	54
8	iobj	29
9	advmod	18
10	nummod	13
11	det	8
12	mark	2
13	vocative	2
14	discourse	1

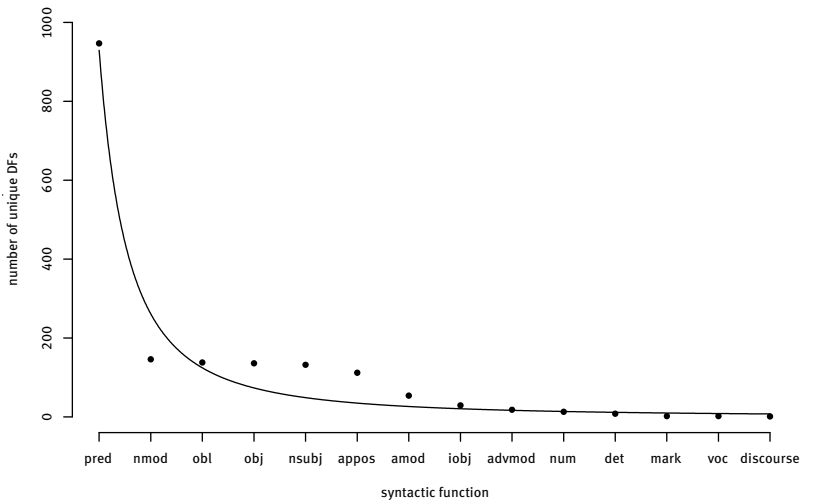


Fig. 8: Number of unique DFs (DF types) for each syntactic function. The line represents function (1) with parameters: $a = 929.42$, $b = -1.8313$.

Tab. 5: Results of fitting function (1) to the data.

Syntactic function	Parameter *a*	Parameter *b*	R^2
pred	1154.18	−0.7603	0.921
nmod	2533.42	−0.9551	0.872
iobj	71.24	−1.0954	0.945
obj	1341.97	−1.1111	0.942
appos	151.25	−1.1513	0.996
nsubj	1805.06	−1.2124	0.966
obl	2903.76	−1.2211	0.969
amod	558.25	−1.3165	0.955
advmod	161.20	−1.3288	0.952
nummod	59.27	−1.4099	0.979
det	41.37	−1.5830	0.993

To get a deeper insight into the analysed phenomena, we observed frequency distributions of DFs for each syntactic function of FEWs as well. We hypothesize that there is a regular frequency distribution of dependency frames for each syntactic function of FEWs (c.f. hypothesis (2) above). To test this hypothesis, we counted ranked frequencies of DFs for each syntactic function separately and then fit function (1) to the data. The results, presented in Tab. 5, show regular ranked frequencies in all cases - the determination coefficient lies in the interval <0.872, 0.996>; this means that hypothesis (2) is not rejected. Further, the coefficient *b* differs with regard to particular syntactic functions. For illustration, the ranked frequencies of syntactic functions with extreme values of *b* are presented in Figures 9 and 10.

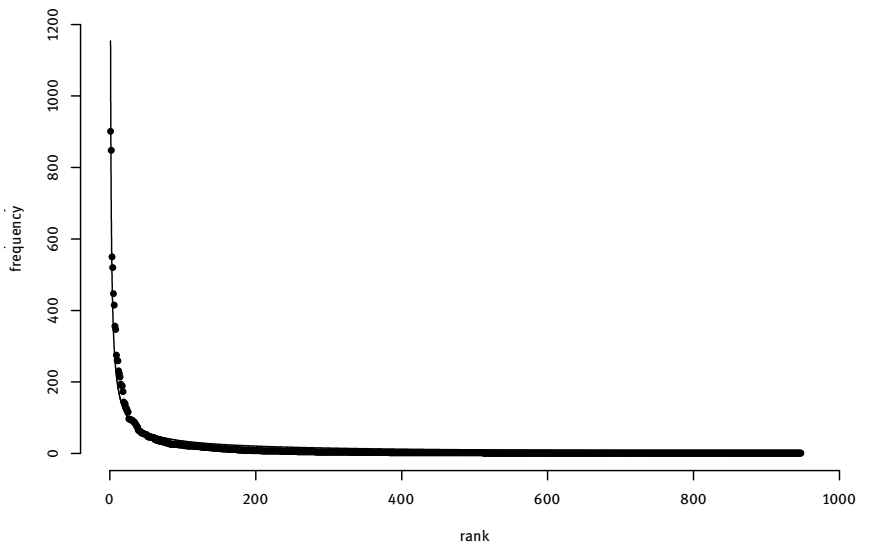


Fig. 9: Ranked frequencies of predicates' DFs (with the highest value of $b = -0.7603$).

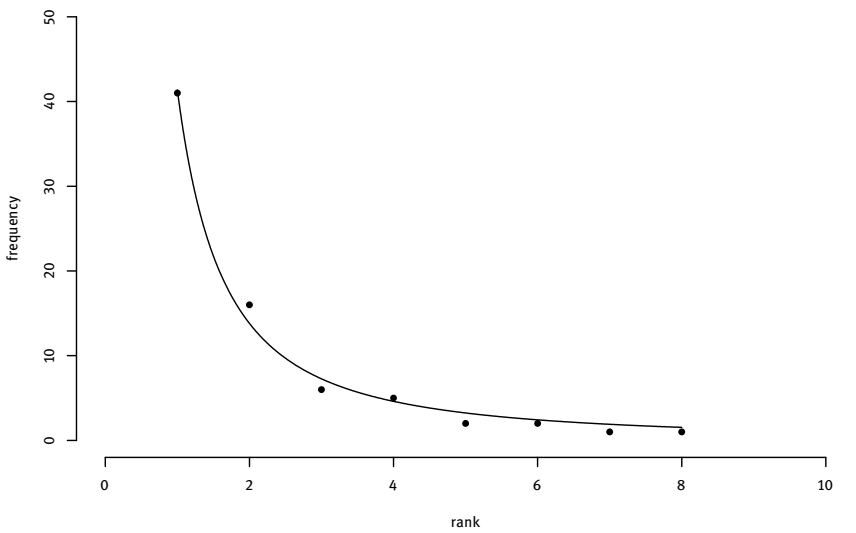


Fig. 10: Ranked frequencies of determinators' DFs (with the lowest value of $b = -1.5830$).

According to the third hypothesis, there should be a relationship between the frequency of FEW with the given syntactic function and the number of unique DFs (DF types) for the given syntactic function. Specifically, we hypothesize: the more frequent the syntactic function, the more unique DFs it has. The hypothesis was tested on the data which are presented in Tab. 6 and Fig. 11.

Tab. 6: Frequencies of FEW with the given syntactic function and of unique DFs (DF types) for the given syntactic function.

Syntactic function	Parameter *a*	Parameter *b*	R^2
pred	1154.18	−0.7603	0.921
nmod	2533.42	−0.9551	0.872
iobj	71.24	−1.0954	0.945
obj	1341.97	−1.1111	0.942
appos	151.25	−1.1513	0.996
nsubj	1805.06	−1.2124	0.966

Syntactic function	Parameter *a*	Parameter *b*	R^2
obl	2903.76	−1.2211	0.969
amod	558.25	−1.3165	0.955
advmod	161.20	−1.3288	0.952
nummod	59.27	−1.4099	0.979
det	41.37	−1.5830	0.993

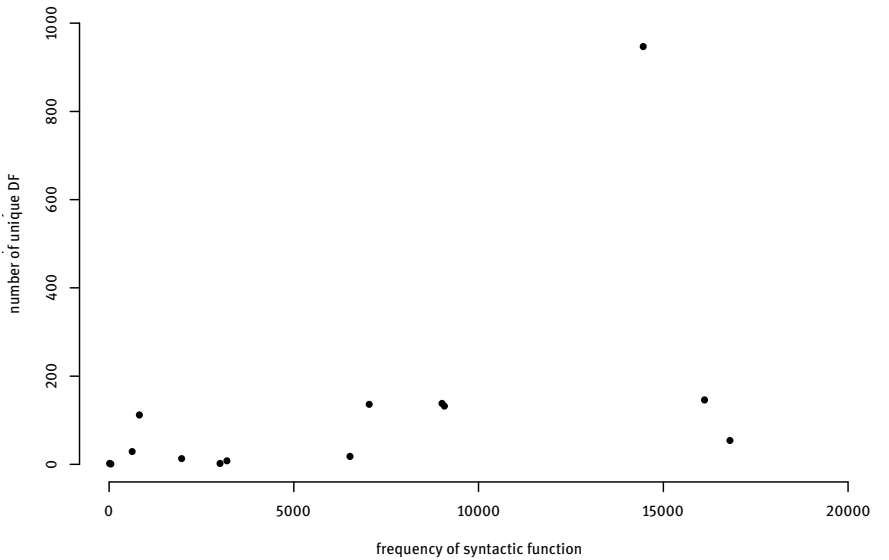


Fig. 11: Frequencies of particular syntactic functions and numbers of their unique DFs.

The hypothesis was tested using the Kendall correlation coefficient which takes the value $\tau = 0.505$. Then the null hypothesis $\tau = 0$ (which corresponds to no increase of the number of unique DFs when the frequency of a syntactic function increases), with the resulting *p*-value of 0.006. Both the relatively high correlation coefficient and the *p*-value do not lead to a rejection of the hypothesis. However, a deeper explanation of the result is needed. First, an extraordinary position of predicate is striking (cf. Fig. 11): its frequent occurrence is accompanied by a very high number of unique DFs which is not, however, a typi-

cal behaviour of other syntactic functions. From the linguistic point of view, this result indicates a high variability of syntactic contexts ruled by the predicate which is connected to its crucial role in a clause. As for the other syntactic functions, there is also positive correlation between frequency of FEW with the given syntactic function and the number of unique DFs. Specifically, if predicate is omitted, the Kendall correlation coefficient is $\tau = 0.529$ (the null hypothesis on the zero correlation is rejected also in this case), which corroborates the hypothesis, too.

5 Conclusion

The results presented in the study reveal three main findings. First, the observed regular frequency distribution of DFs in the corpus can be interpreted as a result of the least effort principle (Zipf 1949) or as an outcome of a diversification process in language (Altmann 2005). It means that this linguistic unit can be included among well-established ones and, consequently, its properties can be used in a general model of language system, such as the model within synergetic linguistic approach.

Second, there are differences among rank-frequency distributions of DFs of syntactic functions of FEWs. The differences are expressed in terms of the values of parameter *b* (see Tab. 5) - the farther from zero, the steeper the curve. Steep curves indicate that there are few dominant DFs which occur very frequently and many DFs with marginal occurrence. On the other hand, the curve for predicate does not decrease so steeply. It seems that frequencies of DFs for predicate are distributed more uniformly than for other syntactic functions which can also be seen in Fig. 11. Predicate is connected to many DFs, but the frequent ones are less dominant (with respect to their frequencies) as for other syntactic functions. The special position of predicate can be - at least partly, as predicate is mostly realized by verbs - explained by a special position of verbs, which was shown by Čech et al. (2011).

Third, there is a relation between the frequency of a syntactic function and the number of its unique DFs, see Fig. 11. For the time being, we are not able to express the relation in terms of a simple mathematical function. The reasons (one of them is a special position of predicate, but there can be many other factors at play, some of which can be stronger than the frequency) present one of challenges for future research.

If the results achieved in this study are corroborated on data from several languages, the regularities observed here can be considered language laws. In

such a case the laws should be incorporated into a language theory and interrelations with other language properties must be established.

It would also be interesting to investigate properties of syntactic dependencies in texts as opposed to corpora. The question whether there are typical parameter values for particular text groups (determined by genres, authors etc.) is more likely to be answered if several syntactically annotated complete texts from the same language are available.

Acknowledgement: J. Mačutek was supported by grant VEGA 2/0054/18.

References

Altmann, Gabriel. 2005. Diversification processes. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp.646–658). Berlin & New York: de Gruyter.

Bejček, Eduard, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek & Šárka Zikánová. 2013. *Prague Dependency Treebank 3.0*. Praha: Univerzita Karlova v Praze, MFF, ÚFAL. http://ufal.mff.cuni.cz/pdt3.0/, Dec 2013.

Čech, Radek, Petr Pajas & Ján Mačutek.2010. Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics, 17*(4), 291–302.

Čech, Radek, Ján Mačutek & Zdeněk Žabokrtský. 2011. The role of syntax in complex networks: local and global importance of verbs in a syntactic dependency network. *Physica A: Statistical Mechanics and its Applications, 390*(20), 3614–3623.

Čech, Radek, Ján Mačutek, Zdeněk Žabokrtský & Aleš Horák. 2017. Polysemy and synonymy in syntactic dependency networks. *Digital Scholarship in the Humanities, 32*(1), 36–49.

de Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC 2006*.

de Marneffe, Marie-Catherine & Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *COLING: Proceedings of the Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Eds.), *Proceedings of LREC 2014*. Paris: European Language Resources Association.

Hudson, Richard. 2007. *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.

Köhler, Reinhard. 1986. *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp.760–775). Berlin & New York: de Gruyter.

Köhler, Reinhard. 2012. *Quantitative Syntax Analysis*. Berlin & Boston: de Gruyter.
Liu, Haitao. 2009. Probability distribution of dependencies based on a Chinese Dependency Treebank. *Journal of Quantitative Linguistics, 16*(3), 256–273.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, 21, 171–193.
Mačutek, Ján, Radek Čech & Jiří Milička. 2017. Menzerath-Altmann law in syntactic dependency structure. In Simonetta Montemagni & Joakim Nivre (Eds.), *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)* (pp.100–107). Linköping: Linköping University Electronic Press.
Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics, 20* (3), 227–240.
Mel'čuk, Igor A. 1988. *Dependency Syntax: Theory and Practice*. Albany, NY: State University of New York Press.
Nivre, Joachim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty & Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk & Stelios Piperidis (Eds.), *Proceedings of LREC 2016*, 1689--1666. Paris: European Language Resources Association.
Wimmer, Gejza & Gabriel Altmann. 2005. Unified Theory of some linguistics laws. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp.791—807). Berlin & New York: de Gruyter.
Zeman, Daniel. 2015. Slavic Languages in Universal Dependencies. In Katarína Gajdošová & Adriána Žáková (Eds.), *Natural Language Processing, Corpus Linguistics, Lexicography* (pp.151–163). Lüdenscheid: RAM-Verlag.
Zipf, George K. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison Wesley.

Anat Ninio

Dissortativity in a Bipartite Network of Dependency Relations and Communicative Functions

Abstract: This paper investigates the relation of syntax and communicative functions. We contrasted Dependency Grammar (DG) that sees syntax as autonomous and independent of its possible communicative uses with Construction Grammar (CxG) that maintains that the building blocks of language are stored form-meaning pairings, including phrasal patterns which may be associated with particular communicative or discourse functions. We constructed two treebanks of Hebrew, of parental speech and of young children's speech, parsing the sentences for DG and coding them for the communicative function of the utterances. The communicative-syntactic system is modeled as a bipartite network of verb-direct object (VO) combinations and of communicative functions (CF). DG predicts that VO-CF matching be scale-free, whereas CxG predicts that the matching is one-to-one. Next, we analyzed the degree of assortativity of the VO-CF bipartite network. We calculated a Pearson's correlation coefficient between the degrees of all pairs of connected nodes, predicting a positive assortativity on CxG and a negative one, on DG. The results do not support the concept of holistic constructions fusing syntactic structures and communicative functions. Instead, the communicative-syntactic system is a complex system with sub-systems mapping onto one another.

Keywords: communicative functions; constructions; bipartite network; assortativity

1 Introduction

The purpose of this paper is to find out how syntax and communicative functions are related in parental speech and in young children's speech. This question has important implications for theoretical linguistics as well as for a theory of language acquisition. Theoretical linguistics in its orthodox version sees syntax as autonomous and independent of its possible communicative uses

Anat Ninio, The Hebrew University, Jerusalem, Israel, Anat.Ninio@huji.ac.il

https://doi.org/10.1515/9783110573565-004

(Chomsky 1988; Saussure 1922/1983), essentially leaving the question open how the one maps onto the other. This conception is adopted by the various versions of the Chomskian tradition as well as many other theories of grammar, among them Dependency Grammar (DG; Tesnière 1959), which is the framework of choice of the present study.

In sheer contrast, a contemporary theory, Construction Grammar (CxG), maintains that the building blocks of language are stored form-meaning pairings, with forms including morphemes and words but also phrasal, clausal or sentence patterns all of which are said to be associated with particular semantic, pragmatic, communicative or discourse functions (Fillmore 1988; Goldberg 2003). In conservative versions of the theory, for example the one developed by Goldberg (1995), syntactic relations are acknowledged, albeit redefined as constructions. That is, some constructions consist of words in a particular syntactic relation such as verb-object, which is then claimed to possess a prototypical meaning with which it is associated. In an extreme version of this theory, Radical Construction Grammar (Croft 2005), it is, however, claimed that there are no syntactic relations such as dependency between words of a sentence, only part-whole relations by which words or sequences of words contribute to the sentence's communicative meaning. This approach rejects formal grammars such as DG on the grounds that they consist of linguistic rules which are merely algebraic procedures for combining words but that do not themselves contribute to meaning, whereas by CxG constructions are meaningful linguistic symbols, being the patterns used in communication. Thus, a linguistic construction is defined as a unit of language that comprises multiple linguistic elements used together for a coherent communicative function, or sub-function such as reference.

To summarize, CxG views the relationship between syntax and communication as a defining association of particular semantic and communicative meanings with certain patterns of surface structure. Which communicative function would be associated with which formal pattern is thought to be determined by the frequency of these patterns in use, hence the so-called "use-conditional" aspect of this theory (Bybee and Hopper 2001).

This theory is particularly influential in the developmental field, giving rise to a model of the acquisition of syntax according to which children do not learn rules for building formal and autonomous structures such as the combination of a verb and a direct object but, rather, they learn the form-meaning pairings of various 'constructions' as the expressions of particular semantic meanings, discourse functions and communicative functions (Ambridge et al. 2015; Goldberg 2003; Lieven 2009; Tomasello 2006). Children's language is believed to be

especially influenced by the prototypicality or the relative frequency of form-function combinations modelled in the parental speech, so that the expectation is that each communicative function would be associated with a single formal pattern, and this association be "preempted" and "entrenched", that is, resistant to change.

Against this model of grammar and of development, in the present paper we raise the hypothesis that sentences are structured with the syntactic relations suggested by DG (Hudson 1984; Tesnière 1959), and that children learn how to produce head-dependent combinations on the basis of the parental speech that serves as input to their learning. According to our hypothesis, head-dependent combinations are the building-blocks of sentences, and the latter are expressions of communicative functions. In this approach, we are embracing the theoretical stance that syntax is autonomous, and that, potentially, any particular head-dependent combination may be used in a sentence serving any communicative function. The detail of how the one is mapped to the other is an open question which we are approaching in the reported study.

The topological character of language that emerges from CxG differs in a crucial manner from that implied by a formal linguistic theory such as DG. Whereas analytic grammars see language as a multileveled complex system, containing separate sub-systems of, among others, syntax and pragmatics (Liu and Cong 2014), CxG sees language as an "inventory of constructions" (Croft 2005; Langacker 1987), each a holistic fusion of form and meaning. In a grammar positing an autonomous syntax, it is an open question how the sub-levels of language are mapped to each other, for example, how are communicative functions mapped to sentences carrying certain syntactic relations, whereas in a grammar positing constructions, form is mapped uniquely to meaning, fused into holistic form-meaning combinations. The overall topological structure of language is accordingly affected. The global structure of a grammar containing an inventory of constructions is a taxonomical network of such fused units, whereas analytic grammars such as DG form multi-leveled complex systems with topological features characteristic of such systems.

These fundamental differences in topological features make it possible to test the two conceptions of grammar against each other. We can ask, are certain syntactic relations uniquely mapped to certain communicative functions and thus form "constructions"? Or, does the relationship of communicative functions and syntactic relations pattern in a manner expected of a complex system?

It is important to realize that not only complexity theory but also theoretical linguistics and philosophy of language on which formal grammars are based, embrace the definition of language as a system. The names most centrally asso-

ciated with this idea are Saussure (1922/1983) and Wittgenstein (1953/1978). Wittgenstein, the philosopher, emphasized in his writings the existence of a complex whole within which individual elements (i.e., particular words or sentences) are meaningful. Language, he pointed out, is similar to a game such as chess, where the existence of the complete game with its set of rules and options is what gives significance to individual moves or pieces. Saussure, the linguist, emphasized the types of connections existing among linguistic units, the interrelations that turn language into a system. His argument for treating language as a system is the dependence of meaning or 'value' of individual words on other words, whether words that occur as their alternatives in sentences, or words that occur together with them, building the structure of sentences. This vision contrasts with CxG's suggestion that language is an "inventory of constructions", namely, a list of units of various sizes; these units may stand in a part-whole relationship with each other but otherwise the inventory is a mere catalogue or collection. These linguistic-philosophical differences appear at first glance to be too abstract to be translated to testable hypotheses. However, with the proper quantitative methods, we can decide between the two conceptions.

We have conducted a study to explore syntax-communications relations both in child-directed adult speech and in young children's speech in order to test the alternative hypotheses regarding the characterization of grammar. We have assembled a large corpus of Hebrew-language spoken utterances produced in the context of parent-child interaction, parsing the sentences for syntax and coding them for the communicative function of the utterances. These data will serve for building two networks representing, respectively, parental and child speech. Our plan is to model the communicative-syntactic system as a bipartite network consisting of two general types of vertices: lexical-syntactic forms, and communicative functions. We shall focus on verb-direct object (VO) combinations and will ask how they form a pattern of connections with the speaker's inventory of communicative functions (CF).

As we have said, the theoretical differences between the two theories of grammar can be expressed as differential predictions regarding the quantitative topological features of the syntax-communication networks. The first difference between the theories touches on the distribution of the connections between VOs and CFs. Networks consist of nodes and the links between them, called edges. The degree of a node in a complex network is the number of incoming or outgoing edges. We shall analyze the distribution of the degrees of specific VO combinations, defined by the verb appearing with any direct object, and of the degrees of specific CFs, encoding the type of talk interchanges the dyads are

engaged in. The two theories of grammar have different quantitative predictions regarding the number of links expected for each node, namely, its degree, as well as regarding the form of the degrees' distribution. According to the hypothesis based on CxG, namely, that language is 'constructions all the way down' (Goldberg 2003), and that children learn specific syntactic constructions as expressions of specific communicative functions (Tomasello 2006), we expect that the VO construction as a whole be strongly associated with some specific type of discourse function. That is, the Constructions approach predicts that VOs and CFs be uniquely connected, as constructions are defined by a unique form-function correspondence. If this is not so for VO in general, then this should be true for individual VOs, namely, combinations of specific verbs with an object. The general linguistic theory predicts that we should find such unique connections in parents' speech, but if not in parents' talk, it should be the case at least in young children's speech as they are supposed to learn such associations and to be "entrenched" in them (Tomasello 2003). In more details, our predictions are as follows.

1) We may expect that VO in general should be associated with some particular CF in maternal speech. We already know that VO is somehow problematic as a construction, as it does not appear to have a prototypical semantic meaning associated with it. The virtual impossibility of matching typical or prototypical semantics to the transitive construction has been extensively discussed in the linguistic literature of English, for instance by Givon (1997), and in the developmental context by Sethuraman and Goodman (2004). As for Hebrew, Glinert (1989) says it quite simply: 'There are no recognized semantic criteria as to which verbs take direct objects.' (p. 159).It is possible though that VO is associated not with a particular semantic prototype but with a specific communicative use. If so, we expect to see this unique association in parental speech. CxG does not specify at which level of generality would a linguistic unit be mapped to a unique CF that provides its meaning. To be exhaustive, we shall test the VO-CF mapping both at the level of an abstract schema of VO and at a lexical-specific definition of individual verbs getting a direct object.
2) Second, it is possible that VO in general is not associated with any specific CF in maternal speech but that individual lexically specific VO combinations are. CxG allows into the grammar both such abstract entities as any VO combination, and lexical-specific entities such as a particular verb with any direct object, for example, "*see something*"

(Goldberg, 1995). Again, we could find that such lexical-syntactic forms are uniquely associated with some specific CF, for example, with requests to see (or not to see) some object.

3) Third, it is possible that even lexical-specific VOs are not associated with a unique CF in maternal speech but, rather, with a set of different CFs. The prediction of CxG in this case is that at least in children's speech, the relevant lexical-specific VOs will be uniquely associated each with a single specific CF, the one most frequently used by mothers with this VO.

In summary, on the basis of CxG we expect that at some level, at the least, VOs and CFs will have unique one-to-one associations.

By contrast, formal grammars such as DG make the opposite prediction regarding the quantitative features of VO-CF combinations. Extrapolating from what we know of the language system under DG assumptions, we expect that VO-couplets and CFs form a joint network that will have the typical topological features characteristic of complex networks. Several studies have shown that networks formed by language units, connected by various different linguistic relationships, exhibit the global statistical features characteristic of complex networks. Naturally occurring complex networks possess some generic topological features, regardless of what are the items making up the network (Baraba'si and Albert 1999; Watts and Strogatz 1998). One of the most important statistical features of complex networks, and possibly their defining attribute, is that the number of links connected to a given node is extremely unevenly distributed. A few nodes have a very large number of links, whereas most nodes have only a very few. For this reason, complex systems are said to be *scale-free,* when the meaning of this term is that there is no typical scale for degrees, and they can range from very low to very high. In previous studies, syntactic, semantic, and phonological networks were found to be scale-free (Ferrer-i-Cancho et al. 2004; Steyvers and Tenenbaum 2005; Vitevitch 2005; Cong and Liu 2014). It is thus expected that the degrees of the various CF and VO nodes should have a scale-free, skewed, power-law distribution. Most certainly it is not expected that the degree distributions would look like a representation of a set of one-to-one mappings.

The further difference between the theories is in their expectation regarding the *assortativity* of the matchings between particular VOs and CFs. Assortativity is the relative tendency of nodes to be connected to other nodes of similar degree. The Constructions approach claims the connection between each particular VO (form) and its particular CF (meaning) is unique, that is, a VO with a

single link is connected to a CF with a single link. The uniqueness of the linking is established and maintained by a couple of principles posited in the theory, *entrenchment* and *preemption* (Tomasello 2006). Entrenchment is a notion developed within Cognitive Linguistics and it is a synonym for learning through repeated exposure or use. This is the major process through which unique mappings of form to function are said to be established and maintained in both adult and child speech. When some form is frequently used for the expression of a specific meaning or function, an "entrenched" association is set up that can inhibit or block the adoption of alternative expressions of that meaning or function (Langacker, 1987). Preemption means blocking the use of alternative expressions of specific communicative intentions if there is in the speaker's system an established (entrenched) form of expression for that intent (Clark and Clark 1979; Goldberg 2005; Pinker 1984). The expression of that function is said to be preempted by the entrenched form of expression, meaning it takes precedence over the potential alternative form. For our purposes, the two hypothesized processes are equivalent, serving to account for the unique form-function correspondence which is the defining characteristic of constructions.

In a second testing of the alternate hypotheses, we shall analyze the degree of assortativity of the VO-CF bipartite network. Assortativity is the relative tendency of nodes to be connected to other nodes of similar degree. If there is no relation between the degrees, the type of connection is ***neutral***. This means that the mapping is random, and the use of any VO for any CF is determined by their relative frequencies. When there is a positive correlation between the connected degrees, we talk about ***assortative matching***. CxG possess two features that predict assortative matching. First, the so-called "use-conditional" aspect of the theory predicts that in any choice situation, highly frequent patterns be chosen over low frequency ones. This predicts that formal patterns such as VO be matched to highly frequent communicative functions, and communicative functions be matched to highly frequent formal patterns, making for a positive correlation. In addition, the posited uniqueness of form-function mapping at the limit translates to the topological prediction that the matching of CFs and VOs is assortative. After competition is resolved, each CF is expected to possess just one VO form that maps to that kind of communicative function, and each VO is expected to be used for one single communicative function. The preference for – and maybe complete restriction to – matching a CF with a single VO-form expressing it, to a VO with a single function expressed, leads to a positive correlation between the degrees of the VO nodes and of the CF nodes.

By contrast, a complex-systems approach predicts ***dissortative matching*** in which the degrees of connected nodes are negatively correlated, so that a

"choosy" VO linked to only a single (or very few) communicative functions, will tend to be the expression of "promiscuous" CF, that is, CF's expressed with many different VOs. Not as CxG, DG does not posit a unique matching of forms to functions. On the contrary, the fundamental autonomy of syntax and communication which is the backbone of an analytic grammar, makes it the norm that forms will be used for multiple communicative purposes, and functions be expressed by many different forms. This makes it perfectly possible that VOs and CFs be "promiscuous" and have large linking degrees. This by itself does not force dissortative matching on the VO-CF connections; that has a crucial role for the system. The CF-VI nodes and links form a complex system, and such system must have complete connectivity. In a scale-free system, the connections must be dissortative in order not to leave nodes with low degrees isolated from the rest of the network. Would the connection be assortative, the network would be fragmented, with "choosy" CFs expressed by "dedicated" VOs not being connected to the rest. This pattern is similar to those observed in other scale-free complex systems. For instance, in the Internet there is dissortative mixing and especially a strong suppression of edges between nodes of low degree of connectivity, which is said to be at least partly attributed to the avoidance of isolated clusters (Maslov, Sneppen and Alon 2003). Indeed in the Internet there are no isolated clusters, and according to our hypothesis, neither are there any in the syntax-communication network.

It is easy to see that CxG not only makes it possible for the form-function network to be fragmented but it actually necessitates this fragmentation. That is, we expect the pattern of connection between VO-combinations and CFs to consist of individually connected VOs and CFs, and, thus, to be fragmented.

We have, therefore, two strongly contrasting sets of hypotheses regarding the assortativity of the VO-CF bipartite network. The prediction derived from CxG is that the matching of CFs to VOs (and the other way) will be assortative, with a positive correlation between the degrees of the linked nodes. In addition, we can predict that there will be in this network many (and maybe most) linked nodes where the degree of the VO node and the CF node are both 1, as the link is unique. By contrast, the hypothesis derived from the Complex Systems approach predicts that the matching will be dissortative, namely there will be a negative correlation between the degrees of the connected nodes. In addition, we predict that there will be no unique matchings of two nodes both with a degree of 1, as this would fragment the complex network. This prediction is in particular for the parental network; it remains to be seen if young children with their limited lexicon manage to build a completely connected network.

An interesting question we may not be able to resolve is which party does the choosing or constraining in the association of forms and communicative uses. It is possible to say that certain forms such as particular VO combinations choose what communicative functions they fill readily, and what functions they will find difficult to serve. For instance, we may think such a VO combinations as *'lost something'* will not easily serve such a function as directing hearer's attention to a new perceptual focus, but will be appropriate for such a function as asking for hearer's help. We might look at such constraints from the other direction and say some communicative functions choose certain language forms such as VO-combinations as appropriate expressions of the relevant function; others will not fit the bill. Verbs and their objects have semantic meaning; the relation between semantics and pragmatics is complex and we would probably be better off treating the selection process as a two-sided one.

As for children's early syntax-communication system, the two theories make clearly contrastive hypotheses. According to developmental theories based on CxG (Goldberg 1995; Tomasello 2006), even if the parental input to development does not contain unique form-function mappings, children would still learn such unique mappings in the shape of "constructions". That is, even if lexical-specific VOs are not associated with a unique CF in maternal speech but, rather, with a set of different CFs, the prediction of CxG is that in children's speech, the relevant lexical-specific VOs will be uniquely associated each with a single specific CF, the one most frequently used by mothers with this VO. The alternative developmental hypothesis based on Dependency Grammar is that syntax is autonomous and the mapping of functions to forms obeys the constraints of a complex system. We expect that children learn not an inventory of isolated form-function couplets but, rather, that they learn language as a complex system. This means that we expect children's bipartite VO-CF network to be similar to mothers'. In prior studies of children's acquisition of syntax it was found that young children's syntactic network is very similar to parents' language network in its global features such as a scale-free distribution (Ninio 2006). Our hypothesis is that children's bipartite network of form-function mapping will also be very similar in its global features to mothers' network, namely, scale-free and dissortative.

Despite its theoretical significance for linguistics and language acquisition theory, there has been no empirical research that systematically investigated the relation between syntax and communicative functions. I believe the reason is mostly methodological: the question is too complex to be dealt with by any but the methods of Complexity Science. Even within Complexity Science, there has been no attempt to represent in the same network both syntax and commu-

nication. There have been a few projects modeling syntactic networks, most notably by Ferrer-i-Cancho et al. (2004) and by Cong and Liu (2014), and some work on semantic networks, for example by Liu (2009), but no attempt yet to model communicative networks, as far as we know. Such a combined language network as the one we are proposing to construct and analyze, will thus constitute an innovation for Complexity Science as well.

The rest of this manuscript will introduce in Section 2 the preparation of two dependency treebanks of spoken Hebrew, and coding for communicative function. Section 3 presents the results of the network analysis and discusses its significance for linguistics and developmental theory. The last section, 4, is the conclusion.

2 Methods and Materials

2.1 Preparation of Two Dependency Treebanks of Spoken Hebrew

In a previous phase of this study, we prepared two dependency treebanks of spoken Hebrew: of adult child-directed speech and of young children's speech. We assembled a large corpus of Hebrew-language spoken utterances produced in the context of parent-child interaction, parsing the sentences for dependency syntax and coding them for the communicative function of the utterances. These data will serve for building two complex networks representing, respectively, parental and child speech.

Speech samples were taken from a videotaped observational study. Forty-eight dyads of mothers interacting with young children acquiring Hebrew as their first language were observed and videotaped in free interaction sessions for 30 minutes at a time in their homes, while engaged in activities of their choosing. The study resulted in 82.5 hours of observations and recorded speech. There were 47 hours of observation of mothers with a high level of education (15 years or more), and 35.5 hours of observation of mothers with a low level of education (up to 10 years). The children in half of each subsample were males, half females, between 10 and 32 months of age, average age about 22 months. The utterances were coded for communicative functions, and parsed for syntactic relations.

Utterances by different mothers in the sample and over different observational periods were pooled. The maternal corpus is considered to give a representative sample of adult speech heard by Hebrew-speaking children at the

relevant age-group. The pooled child corpus is representative of young children's earliest multiword speech.

The corpora of spoken sentences were parsed manually for syntactic structure. We based our dependency analyses on Hudson's Word Grammar (Hudson 1984). We also consulted descriptive grammars of Hebrew such as Glinert (1989) and Ornan (1979).

2.2 Coding for Communicative Functions

All multiword expressions were coded for the type of social-communicative function performed in uttering the utterance, using a detailed category system (Ninio and Wheeler 1987). This system is a reasoned taxonomy of verbal-communicative acts, based on Speech Act Theory (Searle 1969), sociological studies of face-to-face interaction (Goffman 1981), and conversational analysis (Sacks, Schegloff and Jefferson 1974). The taxonomy distinguishes 110 different communicative functions applying to stretches of talk Goffman calls talk interchanges. They fall into families of different types: action negotiation, discussions of joint focus of attention, discussions of nonpresent topics, markings of events, performances of moves in game formats and clarification episodes. The interchanges are further distinguished according to the type of interactive state or event they are related to: for instance, negotiations can be of entering into copresence or of leaving, of getting into focused interaction or of leaving it, of initiating joint action or ending it, of performing single acts or of stopping acts in progress. For example, the taxonomy has five different codes for the initiation of a joint activity, defining the communicative function on a considerable level of detail. The functions distinguished are to initiate a new activity by proposing a specific activity; to initiate a new activity by proposing the performance of a move of that activity; to initiate a new activity by proposing the performance of a preparatory move of that activity; to get hearer to start a new activity after activity has been negotiated; and to initiate a new activity while letting hearer propose the activity. Similar high levels of detail are used for distinguishing other types of functions, e.g., performing moves in different interactive games.

Determination of the communicative function of an utterance was done on the basis of the verbal and nonverbal interactive context of the utterance, as judged from the videotaped observations. Coding was aided by considerations of the participants' nonverbal behaviour, by further clarifications put on the utterance, and by the future course of the conversation.

Coding was done by two, highly trained, coders. On the first run through the data, each coder coded half of the corpora. Five complete observational

sessions were randomly chosen to undergo blind recoding. On the 2,934 utterances in this corpus, the overall inter-coder agreement was 85.1% (*kappa* value 81.1). Subsequent to the reliability check, the data was run through a second time, each coder checking the work of the other. All disagreements were discussed and reconciled.

In this study, only spontaneous utterances were included; we excluded utterances where the function was to imitate a previous utterance, to recite texts of books, poetry or lyrics of songs; or else to complete words and texts on demand. In the maternal corpus there were 61 different communicative functions expressed; in the child corpus, 38.

3 Results and Discussions

The maternal corpus contained 14,036 sentences expressing VO relations. The child corpus consisted of 2,244 sentences with VO.

For statistical analysis, verbs heading VO combinations were classified into verb-stem groups or lemmas by their consonantal root and verb-pattern (*binyan*) value. This analysis discounts tense, gender and plural inflections, and retains the semantic meaning of the verbs. In English this would mean that '*walks*' and '*walk*' are treated as an identical lemma. Tense, gender and number were not seen as relevant for the use of a syntactic combination for communication.

We modelled the communicative-syntactic system as a bipartite network consisting of two general types of vertices (nodes): on the one side, lexical-syntactic forms consisting of VO combinations with specific verbs getting a direct object, and on the other side, the communicative function (CF) of the sentence in which the VO occurred. The links (edges) between the nodes represent speech events in which such an association between form and function occurred. We focused on types of connections and disregarded their token frequency; namely, a link marks the fact that at least one such form-function connection has been observed in the data.

We shall analyze the topological features of the network of connections in the bipartite graph of VO-CF. The relevant measurements are probing the global structural properties of the network. First, we shall analyze the network generated from adult speech, namely, mothers' child-directed speech. This kind of talk is considered the linguistic input for children's acquisition of syntax. Then, we shall analyze children's network, to explore the features of the output of the learning process.

3.1 Form-function Mapping in Maternal Speech

First we tested the alternative form-function hypotheses on adult speech, more particularly, mothers' child-directed speech.

3.1.1 Unique Mapping or Scale-free Mapping

As CxG does not specify at which level of generality a linguistic unit is mapped to a unique CF that provides its meaning, we have tested the hypotheses regarding VO-CF form-function mapping both at the level of an abstract schema of VO and at a lexical-specific definition of individual verbs getting a direct object.

The first possibility derived from CxG is that the multi-word surface structure we call VO is a kind of abstract schema with the features of a construction. As constructions are defined by a unique form-function correspondence, this predicts that adults use the VO construction in general for the expression of some specific type of communicative function. That is, any kind of VO should be uniquely connected to some specific CF, with a degree of 1 in the bipartite network.

Testing this hypothesis, we counted the number of different CFs linked to maternal VOs. The findings were that mothers use the VO pattern for 61 different CFs, not for a single CF. It is obvious that the abstract syntactic pattern of VO as a whole is not uniquely associated with a particular communicative function in the linguistic input. Instead of a single specific CF, mothers use the VO pattern for the maximal number of communicative functions in their repertoire.

The alternative hypothesis derived from DG is that VO as a structural entity be mapped to communicative uses in a scale-free manner. If the prediction of the Construction approach is that the connection of VO and CF is unique, the prediction from DG is that the mapping between the two types of nodes is not characterized by any specific value; the degrees are supposed to be distributed by a function such as power-law.

The second option derived from CxG is that not the complete abstract schema but each specific VO is associated with a unique CF. The bipartite networks constructed had 294 nodes for VOs, 61 nodes for CFs and 1405 edges.

There were 294 different verbs used in the VO pattern in maternal speech. The prediction was that each particular verb occurring with a direct object will be uniquely associated with a single function, that is, its degree in the bipartite network will be 1. According to this hypothesis, there should be at the most 294 different types of VO-CF combinations. In actuality, there were 1,405 different

combination of VO and CF. We computed the degrees of different VO nodes, that is, the number of different communicative functions each served. Fig. 1 presents the distribution of the degrees of specific VOs, in mothers' speech.

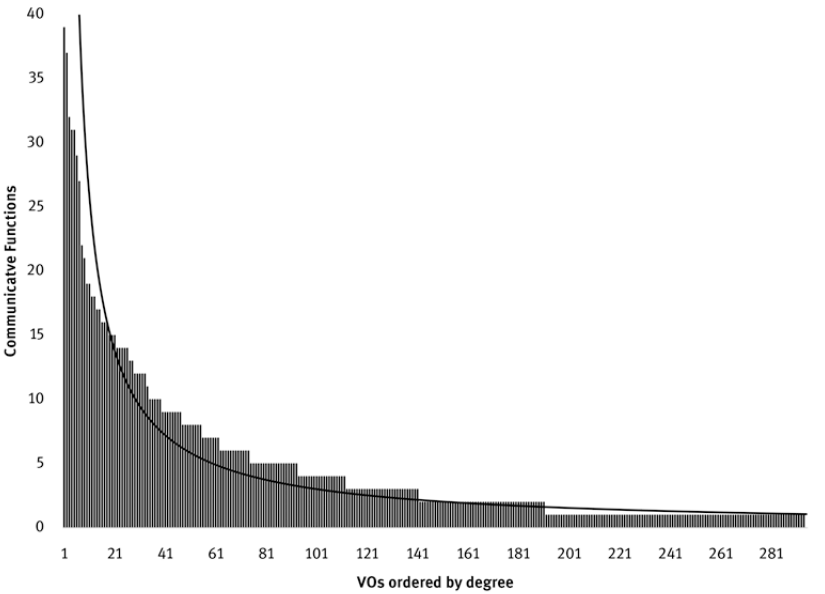


Fig. 1: Distribution of degrees of VO nodes defined by the verb heading the combination, in the bipartite network with CFs, in mothers' speech, with power-law trendline added

We fitted a power-law trendline to the graph, using the option provided by the Excel program. The function was:

$$y = 263.15x^{-0.968} \quad (1)$$

and the fit was R^2= 0.904. That is, the distribution of connectivities of the VO nodes in the bipartite network with CFs is extremely broad, in fact scale-free, and most certainly not unitary.

To summarize, in the first place we tested the radical possibility derived from CxG that all VOs serve a single identifiable CF. This hypothesis was rejected. In the second place, we tested the hypothesis that individual, lexically specific VO combinations – defined by being headed by a particular verb – are uniquely associated each with a single CF in maternal speech. This lexical-specific constructions hypothesis was rejected, too.

3.1.2 Assortativity

In further testing of the alternate form-function mapping hypotheses, we shall analyze the *assortativity* of the VO-CF bipartite network. We shall calculate assortativity by the Pearson's correlation coefficient between the degrees of all pairs of connected nodes (Newman 2002).

There are three options: assortativity is neutral, positive or negative (dissortativity). Neutral assortativity means connections are random and are not influenced by the degree of the node connected to. Positive assortativity means a node with high connectivity will be connected to nodes of high connectivity, and nodes with low connectivity, to nodes of low connectivity. In our case, VOs with many functions (namely, promiscuous VOs) would connect to functions with many VOs (namely, promiscuous functions), and VOs with few functions (namely, choosy ones), to functions with few VOs (namely, choosy ones). If the mapping is one-to-one as expected by the Construction approach, this would mean all VOs and all FCs are choosy, and the assortativity is positive.

In the case of dissortativity, high-connecting VOs would connect to low-connecting functions, and low-connecting VOs, to high-connecting functions. If there are extremely choosy VOs or FCs, with a single linked node of the opposite kind, the prediction based on formal grammars would be that these are linked to promiscuous nodes. The reason is that one-to-one linking would generate fragmentation, with individual linked nodes being separated from the rest of the system. A formal grammar predicts that the bipartite network of VOs and CFs is a complex system, namely, it is totally connected. That implies dissortativity.

There were 1,405 different combinations of CFs and VOs. We computed a Pearson correlation coefficient between degrees of linked CF and VO nodes; the coefficient was −0.43 with 1,403 degrees of freedom. This value is significant at the $p < 0.001$ level.

That is, in mothers' network of form-function mapping, we found dissortativity (negative degree correlations), and it is as predicted by a formal grammar such as Dependency Grammar, not as predicted by Construction Grammar. There are many different possible explanations for dissortativity (for instance, see Maslov and Sneppen 2002). We tied the possible dissortativity of the network to the suppression of low-connectivity nodes with one link avoiding links to low-connectivity nodes of the opposite type, also with one link. We derived the prediction that the bipartite network of syntax and communication be dissortative, and unique mapping be avoided, from the need of language networks to be all-connected. A similar explanation for the strong suppression of connections between pairs of nodes of low connectivity was offered for the dissortativi-

ty of the Internet (Maslov, Sneppen and Zaliznyak 2004), as the Internet operates under the constraint that clusters of autonomous systems do not stay isolated from each other but have to be connected to other parts of the net by at least one path. In a less obvious way, language also operates under the constraint that it be completely connected. We heard from Wittgenstein (1953/1978) and Saussure (1922/1983) that language is a system and that meaningfulness is achieved by contrast. We can operationalize this very general concept to the requirement associated with complex systems, namely complete connectivity.

To test if this explanation is correct, we checked if indeed there was complete connectivity in the bipartite network connecting the syntactic pattern VO to communicative uses. To achieve connectivity in a bipartite network, there cannot be isolated couplets where a given VO is linked to only a single CF, and that CF is only linked to that VO. Checking the data, there were 103 VO types with only a single CF, but in none of the cases was that function used only with the relevant "choosy" VO. Similarly, there were 11 different CFs (18.0% of the inventory) that were expressed by only a single VO, but all these had also several other functions. In the whole network, there was no one-to-one mapping at all, thus no isolated VO-CF units. The network has complete connectivity.

It appears that in the syntax-communication system of adult child-directed speech, there is a suppression of connections between nodes with low connectivity. Take note that the expected pattern according to Construction Grammar is one-to-one mapping which we did not find at all in our data. It may be summarized that the maternal network of form-function connections involving the VO pattern does not appear to be composed of constructions, but, rather, of autonomous forms used for a variety of functions, as long as this use does not violate the constraint that the whole network be connected.

We turn now to young children's early syntax-communication system. The hypothesis derived from CxG expect this system to consist of unique form-function mappings, whereas the hypothesis derived from DG expects children's bipartite VO-CF network to be a complex system and similar to mothers'.

3.2 Form-function Mapping in Young Children's Speech

Children use the VO pattern with 38 different CFs, rather than with a single CF. This rejects the extreme hypothesis derived from CxG that the multiword surface structure we call VO is a kind of abstract schema with the features of a construction. As constructions are defined by a unique form-function correspondence,

this predicts that children use the VO construction in general for the expression of some specific type of communicative function. Apparently, VO is not an abstract schema/construction for young children.

Moving on to the second option derived from CxG, each specific VO may be associated with a unique CF. The bipartite networks constructed for children had 127 nodes for VOs, 38 nodes for CFs and 465 edges.

There were 127 different verbs used in the VO pattern in child speech. The prediction was that each particular verb occurring with a direct object will be uniquely associated with a single function, that is, its degree in the bipartite network will be 1. According to this hypothesis, there should be at the most 127 different types of VO-CF combinations. In actuality, there were 465 different combination of VO and CF. We computed the degrees of different VO nodes, that is, the number of different communicative functions each served. Fig. 2 presents the distribution of the degrees of specific VOs, in children's speech.

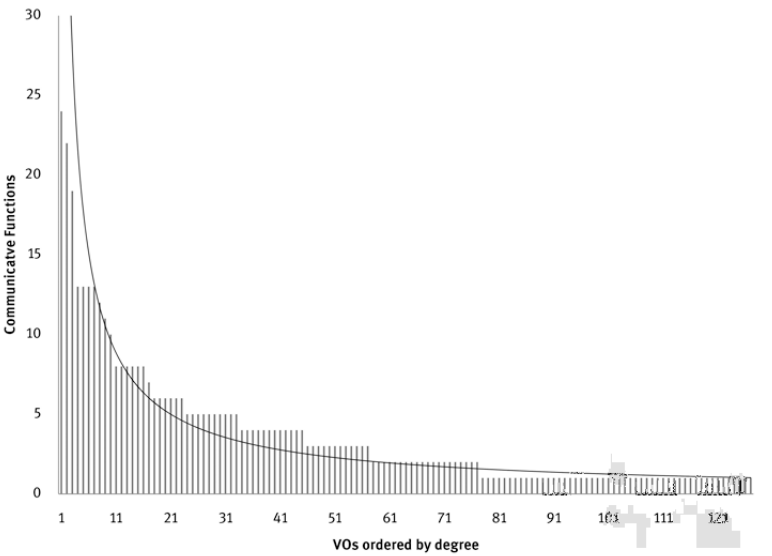


Fig. 2: Distribution of degrees of VO nodes defined by the verb heading the combination, in the bipartite network with CFs, in children' speech, with power-law trendline added

We fitted a power-law trendline to the graph. The function was:

$$y = 73.702x^{-0.884} \tag{2}$$

and the fit was R^2= 0.9046. That is, the distribution of connectivities of the VO nodes in the bipartite network with CFs is scale-free, and not unitary.

The slope of the power law function and its fit are practically identical to mothers'. The similarity of the two graphs is striking, especially if we remember that children have less that half number of different verbs in VO than mothers, and about half as many CFs.

Next, we measured the assortativity in children's network. There were 465 different types of combination of VOs and CFs. The Pearson's correlation coefficient between the degrees of all pairs of connected nodes (Newman 2002) was -0.38 (463 D.F., p<0.001), very similar to mothers'.

As in mothers' network, the negative assortativity is connected to avoidance of isolated clusters. There were 50 VO types with only a single CF, but in none of the cases was that function used only with the relevant "choosy" VO. Similarly, there were 10 different CFs (26.3% of the inventory) that were expressed by only a single VO, but all these VOs had additional functions. In the whole network, there was no one-to-one mapping at all, thus no isolated VO-CF units. The child network has complete connectivity.

4 Conclusions

Theoretical issues in linguistics such as the relative validity of Dependency Grammar versus Construction Grammar are usually not approached with the tools of the quantitative analysis of syntactic networks. This, however, was the goal of the present paper. By using the tools of statistical physics of complex networks, we validated the use of Dependency Grammar to describe syntactic patterns and showed that Construction Grammar fails to predict the topological features of the syntax-communication network. We also showed that maternal and child syntactic-communicative dependency networks share their major quantitative features such as scale-free degree distributions and dissortative matching of syntactic schemas to communicative functions. We did not find evidence for "entrenchment", "preemption" or any other process that would generate the one-to-one mapping of syntax and pragmatics expected from constructions. The relation of syntax and communication does not consist of a list with entrenched associations between VO-phrases and certain CFs. Instead, we found that both mothers' and children's syntax and pragmatics form a complex system with typical characteristics. First, VO as an abstract syntactic construction does not get used for the expression of a single type of communicative function but, rather, for the expression of many different types of functions. Second,

when we look at specific VO phrases defined by the use of particular verbs, we find that syntax and pragmatics form a network with dissortative matching, mainly because of a significant difference in the numbers of nodes for VOs and CF. In such conditions, dissociative matching is the only way to achieve a completely connected network. Most significantly, the considerable difference in the numbers of specific nodes for VOs and CFs is part of the findings, namely, that the mapping between syntax and pragmatics is not one-to-one but a complex many-to-one and one-to-many mapping between syntactic types and communicative types. We have also shown that children learn the global features of the system very early on; what is left to learn is more of the concrete linguistic items constructing the system: young children possess fewer verbs in the VO pattern and fewer CFs than there are in mothers' speech.

One alternative we need yet to discuss concerns the possibility that VOs fulfill not complete communicative functions but, instead, sub-functions which are components of many different communicative functions. Tomasello (2006) gives NP as an example of a phrase filling such a sub-function, pointing out that, regardless of its specific composition, a NP always serves to make reference. Although this claim is not quite correct (there are NPs that serve as predicate complements without referencing any specific entity), it is worth examining in more detail. Tomasello gives NPs as an example of a phrase serving a subfunction with a wide spread in different communicative functions, with other types of phrases evidently behaving similarly and filling other kinds of subfunctions. The question is, if NPs fill sub-functions, why not VOs also? The answer is that, besides VO, there are many different verb-based argument structures, for instance verb-indirect-object, verb-prepositional object, verb-adjunct, verb-infinitive, verb-gerund, verb-that-clause, verb-predicate complement, and combinations of them (as we can see in Hornby 1945). Assuming that, like we found in this study for VO, all these verb-based patterns serve many different CFs, according to the sub-function theory each should have some communicative sub-function such as reference that can apply to many different communicative contexts. There is however a grave problem with the theoretical foundations of this possibility. We know from Searle's (1969) seminal work on speech acts that there are only two sub-functions of speech acts, which Searle calls prepositional acts: reference and predication. As far as philosophy of language is concerned, there are no other sub-functions that may be taken up by surface strings that we call phrases, and in fact it is difficult to think of any others that will hold water. It is reasonable that NPs (or DPs in present theories) are often used for reference, but it is less plausible that we can attach some meaningful function to such a unit as verb-object combinations. Until researchers in the

Radical Construction Grammar tradition come up with a robust new Speech Act Theory replacing Searle's in which there are many well-argued sub-functions, it seems that this idea will not be tenable.

Acknowledgement: This research was supported by the United States-Israel Binational Science Foundation (BSF), Jerusalem, Israel, to Anat Ninio and Carol Eckerman under Grant no. 2467/81, by The Spencer Foundation, US under Grant no. 200900206, by The Center for Complexity Science (CCS) to Anat Ninio and Sorin Solomon, under Grant no. GR2007-043, and the Israel Foundations Trustees (Ford Foundation) under Grants no. 35/92 and 13/94.

References

Akhtar, Nameera & Katherine H. Herold. 2009. Pragmatic development. In Marshall Haith & Janette B. Benson (Eds.), *Encyclopedia of Infant and Early Childhood Development,* Vol. *2* (pp. 572–581). San Diego, CA: Academic Press.

Ambridge, Ben, Amy Bidgood, Katherine E. Twomey, Julian M. Pine, Caroline F. Rowland & Daniel Freudenthal. 2015. Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE, 10* (4), e0123723.

Baraba'si, Albert–László & Réka Albert. 1999. Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Bybee, Joan L. & Paul J. Hopper. (Eds.). 2001. *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins.

Chomsky, Noam. 1988. *Language and Problems of Knowledge: The Managua Lectures*. Cambridge, MA: The MIT Press.

Clark, Eve V. & Herbert H. Clark. 1979. When nouns surface as verbs. *Language, 55*(4), 767–811.

Cong, Jin & Haitao Liu. 2014. Approaching human language with complex networks. *Physics of Life Reviews, 11,* 598–618

Croft, William A. 2005. Logical and typological arguments for radical construction grammar. In Jan–Ola Östman & Mirjam Fried (Eds.), *Construction Grammars: Cognitive Grounding and Theoretical Extensions* (pp. 273–314). Amsterdam: John Benjamins.

Ferrer–i–Cancho, Ramon, Ricard V. Solé & Reinhard Kohler. 2003. Universality in syntactic dependency networks. *Santa Fe Institute Working Paper #03–06–042*.

Fillmore, Charles J. 1988. The mechanisms of "Construction Grammar". *Proceedings of the Fourteenth Annual Meeting of the Berkeley Linguistics Society* (pp. 35–55).

Givon, Talmy. 1997. Grammatical relations: An introduction. In Talmy Givon (Ed.), *Grammatical Relations: A Functionalist Perspective* (pp. 1–84). Amsterdam: John Benjamins.

Glinert, Lewis. 1989. *The Grammar of Modern Hebrew*. Cambridge: Cambridge University Press.

Goffman, Erving. 1981. *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Goldberg, Adele E. 1995. *Construction Grammar*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences, 7* (5), 219–224.

Goldberg, Adele E. 2005. Argument realization: The role of constructions, lexical semantics and discourse factors. In Jan–Ola Östman & Mirjam Fried (Eds.), *Construction Grammars: Cognitive Grounding and Theoretical Extensions*. (pp. 17–43). Amsterdam: John Benjamins.

Hornby, Albert Sydney. 1945. *A Guide to Patterns and Usage in English*. London: Oxford University Press.

Hudson, Richard. 1984. *Word Grammar*. Oxford: Basil Blackwell.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar, Vol. 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.

Lieven, Elena. 2009. Developing constructions. *Cognitive Linguistics, 20* (1), 191–199.

Liu, Haitao. 2009. Statistical properties of Chinese semantic networks. *Chinese Science Bulletin, 54* (16), 2781–2785.

Liu, Haitao & Jin Cong. 2014. Empirical characterization of modern Chinese as a multi-level system from the complex network approach. *Journal of Chinese Linguistics, 42* (1), 1–38.

Maslov, Sergei & Kim Sneppen. 2002. Specificity and stability in typology of protein networks. *Science, 296*(5569), 910–913.

Maslov, Sergei, Kim Sneppen & Uri Alon. 2003. Correlation profiles and motifs in complex networks. In Stefan Bornholdt & Hans Georg Schuster (Eds.) *Handbook of Graphs and Networks: From the Genome to the Internet* (pp. 168–198). Berlin: Wiley-VCH.

Maslov, Sergei, Kim Sneppen & Alexei Zaliznyak. 2004. Detection of topological patterns in complex networks: Correlation profile of the Internet. *Physica A: Statistical Mechanics and its Applications, 333,* 529–540.

Newman, Mark E. J. 2002. Assortative mixing in networks. *Physical Review Letters, 89*, 208701–208704.

Ninio, Anat. 2006. *Language and the Learning Curve: A New Theory of Syntactic Development*. Oxford: Oxford University Press.

Ninio, Anat & Polly Wheeler. 1987. A manual for classifying verbal communicative acts in mother–infant interaction-revised. *Transcript Analysis, 3*(1), 1–83.

Ornan, Uzi. 1979. *Hamishpat Hapashut [The Simple Sentence]*. Jerusalem: Academon.

Pinker, Stephen. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

Sacks, Harvey, Emanuel A. Schegloff & Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Saussure, Ferdinand de. 1922/1983. *Course in General Linguistics*. (R. Harris, Trans.). London: Duckworth.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

Sethuraman, Nitya & Judith C. Goodman. 2004. Children's mastery of the transitive construction. In Eve V. Clark (Ed.), *Online Proceedings of the 32nd Session of the Stanford Child Language Research Forum* (pp. 60–67). Stanford, CA : CSLI Publications.

Steyvers, Mark & Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*(1), 41–78.

Tesnière, Lucien. 1959. *Éléments de Syntaxe Structurale*. Paris: Klincksieck.

Tomasello, Michael. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, Michael. 2006. Construction grammar for kids. *Constructions*, Special Volume 1, 11/2006.
Vitevitch, Michael S. 2005. *Phonological Neighbors in A Small World: What Can Graph Theory Tell us about Word Learning?* Paper presented to the Complex Systems and Networks Group at Indiana University.
Watts, Duncan J. & Steven H. Strogatz. 1998. Collective dynamics of 'small–world' networks. *Nature, 393*, 440–442.
Wittgenstein, Ludwig. 1953/1978. *Philosophical Investigations* (Gertrude E. M. Anscombe, Trans.). Oxford: Blackwell.

Andrei Beliankou, Reinhard Köhler*

Empirical Analyses of Valency Structures

Abstract: Verb valency structures have long been the focus in dependency grammar. Arguments, whether they are obligatory complements or optional adjuncts of verb valency, generally are not described with respect to conditions and criteria in linguistic dictionaries. Therefore, this article attempts to examine the individual dynamic occurrence of verbs with their specific argument structures, and the differentiation between complements and adjuncts from a quantitative perspective. Data were obtained from a dependency syntactically annotated Russian National Corpus (RNC), which contained 2255 different Sentence Structure Schemes (SSS). Several conclusions are obtained. First, the rank-size distribution of SSS fit the Waring distribution perfectly. Second, the relation between the number of SSS (y) and the number of verbs (x) with y SSS follows an exponential function. Third, the number of complements for SSS and the frequency of the SSS with x complements abide by the Binomial distribution. The empirical data concerning valency structures display a lawful behaviour, which has relative implications for language teaching and learning.

Keywords: verb valency structures; distributions; Sentence Structure Scheme (SSS)

The traditional paradigm focuses on the verb (valency can be attributed not only to verbs but also to other parts of speech such as nouns and adjectives) as the centre of a sentence and starts with the assumption that the arguments of the verb are either obligatory complements or optional adjuncts (Tesnière 1959; Comrie 1993; Heringer 1993). However, so far it has not been possible to give satisfactory criteria for this distinction. If a verb which is described as calling for an obligatory complement but is found in natural texts without any, proponents often call the correspon-ding expression elliptic. And a verb with more than the mandatory dependents has, so they say, just optional dependents. General conditions and criteria are missing such that the situation is handled only with the help of names, as often in linguistics. For a first attempt at a quantitative analysis of valency phenomena, an approach had to suffice which relied on an estab-

Andrei Beliankou, Universität Trier, Trier, Germany
Reinhard Köhler, Universität Trier, Trier, Germany, koehler@uni-trier.de

https://doi.org/10.1515/9783110573565-005

lished valency dictionary. In Köhler (2005 and 2012, 92–113), the well-known dictionary of verb valency by Helbig and Schenkel (1991) was consulted, which covered a large number of details on 500 frequent German verbs.

The dictionary provides for each of the verbs (a) verb variants, which differ in valency and meaning, and was shown to abide by a diversification model, viz. the positive negative distribution.

Each of the verbs is attributed (b) one or more sentence structure scheme ("Satzbauplan"), coded by Köhler in a computer-friendly but still easily readable way, e.g.: The code

```
Sn (pSa/Ad j) Part/Inf
```

describes a sentence pattern with a subject in nominative case facultatively followed by either a prepositional object in accusative case or an adjective, and an obligatory infinitive or a participle. The frequency distribution can be modelled using the Zipf-Mandelbrot distribution with extremely good results.

Another very interesting detail is the semantic sub-categorisation Helbig and Schenkel provide (c): each argument is sub-specified in terms of lexical-semantic categories such as abstract, human, abstract-human (e.g., university, police), +anim/-anim, -indiv (excluded individuals), action. This information is given for each argument and enables the researcher to consider the distribution of the number of alternative semantic categories a verb (variant) accepts. An appropriate model is the Poisson distribution as another diversification distribution.

In his publications (2005 and 2012, 101ff), Köhler addresses also the problem of this kind of valency dictionaries which are based on the description of verbs and other parts of speech by linguists without any visible empirical source – a problem which does not occur when an empirical and quantitative approach to valency is combined with theoretical considerations. As opposed to a pure dictionary-based approach where valency is viewed as a constant property of words, another point of view is possible: We can for instance determine the arguments of the verbs in a corpus, i.e., observe the individual occurrences of verbs with their specific argument structure. In this way, the differentiation between complements and adjuncts can be considered as a gradual or quantitative criterion – or it can be abolished and replaced by a quantitative property, viz. valency as a tendency to bind other words.

Moreover, if valency is defined in a way such that not only the number of dependents or arguments is of interest but also the type, we can, after determining and annotating the dependency types of the individual dependents of a

head, study the distribution of these link types, too. The Russian corpus[1] used in Köhler (2012) provides this information: the links from the heads to the dependents are categorised; each link token in the corpus was assigned one of the types. The corpus differentiates the 63 dependency types listed in Table 3.25 on page 102 in this publication. Other authors published also valency analyses based on corpus data, cf. e.g. Čech, Radek; Pajas, Petr; Mačutek, Ján (2010).

Here, we studied Russian dependency with data obtained from the syntactically annotated part of the Russian National Corpus (Apresjan et al. 2005; Plungian 2009). This corpus is an electronic reference system of modern Russian and provides morpho-syntactic and semantic annotations on the basis of the "Meaning & Text" model by Igor A. Mel'čuk and Alexander K. Zholkovsky. The corpus provides both spoken (transcribed) and written texts. The RNC consists primarily of original prose text but also includes smaller amounts of translated works. As for the current state, the corpus contains about 150 millions of annotated data tokens.

The material contains 2255 different Sentence Structure Schemes (SSS). Tab. 1 shows the head and the tail of the frequency-sorted list and Tab. 2 further explains the tags used in Tab. 1.

Tab. 1: List of the most frequent and some of the least frequent SSS

SSS	Frequency	SSS	Frequency
Empty SSS	6128	"ADV"	1610
"PR"	4848	"S_ТВОР"	1478
"S_ВИН"	4795	...	...
"PR S_ИМ"	3679	...	...
"S_ИМ V"	2506	"S_ИМ S_РОД S_РОД S_ТВОР"	1
"S_ИМ"	2380	"S_ИМ S_РОД S_РОД V"	1
"PR S_ВИН"	2267	"S_ИМ S_РОД S_ТВОР S_ТВОР V"	1
"V"	1918	"S_ИМ S_ТВОР S_ТВОР V"	1
"S_ВИН S_ИМ"	1791	"S_ИМ V V V V"	1

1 SYNTAGRUS: A Corpus of Russian Texts Syntactically Annotated with Dependency Trees, developed by the Laboratory of Computational Linguistics of the Institute for Problems of information Transfer of the Russian Academy of Sciences, Moscow.

Tab. 2: The explanation of tags used in Tab. 1

Tag	Description	Tag	Description
A	Adjective	S_NOCASE	Noun without case marker
ADV	Adverb	S_ВИН	Noun (Accusative)
COM	Composed Expression	S_ДАТ	Noun (Dative)
CONJ	Conjunction	S_ЗВ	Noun (Vocative)
INTJ	Interjection	S_ИМ	Noun (Nominative)
NID	Foreign Expression (Non-Russian)	S_ПАРТ	Noun (Partitive)
NUM	Numeral	S_ПР	Noun (Prepositional)
P	Pronoun	S_РОД	Noun (Genetive)
PART	Particle	S_ТВОР	Noun (Instrumental)
PR	Preposition	V	Verb

As can be seen, on top of the list there is an SSS without any complement, the tail is formed of hapax-legomena tokens. The empirical distribution resembles a typical rank-size distribution, as known from word rank-size distributions and distributions of other elementary units in linguistics. Therefore, a corresponding probability distribution can be expected. We attempted to fit the Waring distribution to the data and obtained a C coefficient[2] of 0.0199 with 2251 degrees of freedom and sample size 87327. This is an extraordinarily good value given the sample size and the number of degrees of freedom. This result could be interpreted as an argument for the status of SSS as linguistic units and support the view advanced in Köhler (1999).

Next, we investigated the number x of verbs with y SSS according to the data from the corpus. The extreme cases are one verb with 617 different SSS. It is the verb “est”, the Russian copula “be”. On the other hand, there are 2079 verbs which allow only a single SSS. The rest of the values follow a skew line.

We assume that the diachronic process whose long-term result in the current situation can be modelled by a diversification approach. A (only hypothetical) start configuration is a small number of verbs with only one SSS each. New verbs which come as neologisms may take over one of the existing SSS *per analogiam*; such that more and more verbs share an SSS. We also assume that the number of such verbs decreases monotonously following an exponential function. We can start the modelling using Altmann's (1980) approach which he

2 C is a function of Chi-square, which is used in case of a huge number of data.

set up to derive a function as a model of the relation between length of a linguistic construct and the lengths of the components of that construct. The resulting function was baptised “Menzerath’s Law” and is now called “Menzerath-Altmann Law”. Here, we will set up a model of the growth of the number of verbs which use an SSS. In the differential equation,

where y denotes the number of verbs, y' its first derivative, and $-b$ is the (negative) growth rate.

$$\frac{y'}{y} = -b \tag{1}$$

By integration, we obtain the solution to this differential equation – the function

$$y = ae^{-bx} \tag{2}$$

Here, a is the integration constant. The value of this parameter will always be unity or less because a set of SSS without any verb using it does not make any sense and cannot be found in the data, of course. We add a term "+1" to the function to exclude this impossible case.

The number of verbs with y Structure Schemes, or frames, can thus be calculated by using the formula.

$$y = 1 + ae^{-bx} \tag{3}$$

The result of the fit using the optimising program NLREG is very good. Parameter estimation yielded

a = 3909.4804 and

b = 0.6634

and the coefficient of determination yields the value

R^2 = 0.9795.

These findings lead to another research question and another study: The SSS have different numbers of complements, from no complements[3] up to 6128. We will now ask for the distribution of the sizes of the SSS in our material. We can assume that the number of SSS with size x depends on the number of SSS of size x—1. The idea of such an assumption is that in a hypothetical early stage of a language, the first formed SSS were short and had no or only one complement.

3 In Russian, complement-less verbs are grammatical as opposed to e.g. German.

Later, when communication need required more complex expressions, new sentence structures were invented or admitted, mainly by expanding some of the existing SSS. Therefore, the amount of more complex structures has to be proportional to the number of previously existing ones. The following factors must play a role in this dynamic development:

a. the need for more complex expressions, i.e. for SSS with more complements than x
b. the wish to avoid complexity because it is connected to effort
c. existing complexity and effort at a given state

Hence, we will assume an increase in complexity for some time, followed by a turning point when effort becomes too large to be compensated by the benefit of complexity. We have, however, no data on the historical development but we may be sure that, after a long time, a synchronic state will mirror this development. We think that a language keeps a kind of equilibrium as long as no drastic events disturb it.

In mathematical terms, our hypothesis is that there is a constant need to increase the potential complexity of sentence structures b, and another constant reflecting the need to save effort a. Then, let us assume that b will become smaller the higher the current complexity is. This is in form of a mathematical expression

$$\left(\frac{b}{x} + a\right) \tag{4}$$

With the help of Altmann's (1991) approach, we can determine the probabilities P_x of the

$$P_x = \left(\frac{b}{x} + a\right) P_{x+1} \tag{5}$$

sizes x recursively:

This formula can be transformed into the usual form of a probability distribution

$$P_x = \binom{n}{x} p^x q^{n-x} \tag{6}$$

which is known as Binomial distribution.

We can now test this hypothesis on data from the Russian corpus using again the iterative optimisation algorithm in the Altmann-Fitter and the fitting

result is presented in Tab.3. The result is positive as can be seen from the numerical results and also from the diagram. The value of C is extremely small, which indicates a perfect fit. The estimated parameters are n=8 (the value of x_{max}), and p= 0.2783, which is, by the way, almost the value of the mean of the empirical distribution $\bar{x}$= 2.2256. Cf. also Fig. 1, which shows clearly that the theoretical and empirical bars are very close to each other.

Tab. 3: Fitting result of SSS based on Russian National Corpus

x[i]	f[i]	NP[i]
0	6128	6428.24
1	20402	19828.87
2	26394	26759.77
3	20996	20636.15
4	9706	9946.15
5	2960	3068.04
6	643	591.49
7	88	65.16
8	10	3.14

Notes: x[i] represents the class number, f[i] represents the observed value and NP[i] is the theoretically expected value

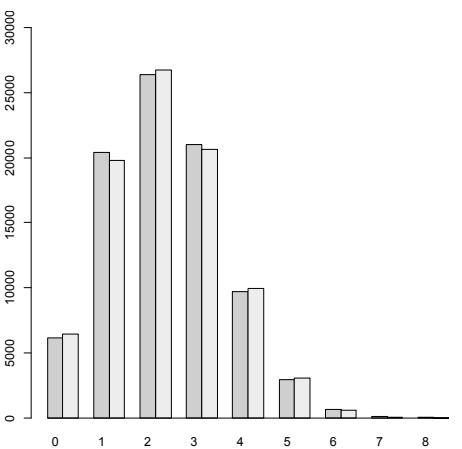


Fig. 1: Graph of the Binomial distribution as fitted to the data from Russian

We conclude from our little study that empirical data concerning valency structures show a lawful behaviour. The distinction between complements and adjuncts can be considered artificial and may help as a means for language teaching and learning but does not add any testable linguistic knowledge. Data which describe verbs according to these categories lead to different parameters in quantitative analyses but do not seem to contradict our and other researchers' approaches. Many more studies will be needed for a general model of the various aspects of valency, in particular data from as many languages as possible.

References

Apresjan, Yu et al. 2005. Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л. и др. Синтаксически исемантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка: 2003—2005. М.:Индрик, 2005, 193—214.

Altmann, Gabriel. 1980. Prolegomena to Menzerath's Law. *Glottometrika, 2*, 1—10.

Altmann, Gabriel. 1991. Modelling diversification phenomena in language". In Ursula Rothe (Ed.), *Diversification Processes in Language: Grammar* (pp.33—46). Hagen: Margit Rottmann Medienverlag.

Čech, Radek, Petr Pajas & Ján Mačutek. 2010. Full valency: Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics, 17*(4), 291—302.

Comrie, Bernard. 1993. Argument structure. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (Eds.), *Syntax: Ein internationales Handbuch zeitgenössischer Forschung. Halbband 1* (pp.905—914). Berlin/New York: de Gruyter.

Helbig, Gerhard & Wolfgang Schenkel. 1991. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Tübingen: Max Niemeyer Verlag.

Heringer, Hans Jürgen. 1993. Basic ideas and the classical model. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (Eds.), *Syntax: Ein internationales Handbuch zeitgenössischer Forschung. Halbband 1* (pp.293—316). Berlin/New York: de Gruyter.

Köhler, Reinhard. 1999. Syntactic structures: Properties and Interrelations. *Journal of Quantitative Linguistics, 6*(1), 46—57.

Köhler, Reinhard. 2005. Quantitative Untersuchungen zur Valenz deutscher Verben. *Glottometrics*, 9, 13—20.

Köhler, Reinhard. 2012. *Quantitative Syntax Analysis*. Berlin/Boston: De Gruyter Mouton.

Plungian V.A. (Ed.). 2009. *Nacionalnyj korpus russkogo jasyka: 2006-2009*. Novyje rezultaty i perspektivy. SPb.: Nestor-Istorija.

Tesnière, Lucien. 1959. *Éléments de Syntaxe Structural*. Paris: Klincksieck.

Huiyuan Jin, Haitao Liu*

Regular Dynamic Patterns of Verbal Valency Ellipsis in Modern Spoken Chinese

Abstract: Valency is an important notion in dependency grammar. This research attempts to observe verb valency patterns from a dynamic perspective. Ellipsis provides us with a good opportunity to observe the regular dynamic patterns of verbal valencies. Three major conclusions are obtained: first, the rank-frequency distributions of ellipsis-related verbs with different valencies obey power law; second, the distributions of the elliptical patterns for the three categories of verbs obeys power law; verbs with different valencies display similar elliptical patterns to some degree, yet discrepancies still exist; third, the frequency of elliptical valency for two-valency verbs and the frequency of the dependency relations of the elliptical constituents in the three categories fit the power law. Moreover, the ellipsis of obligatory constituents are more likely to be subjects syntactically, or agents from a semantic aspect.

Keywords: valency; verb; ellipsis; rank-frequency distribution

1 Introduction

The notion of valency was introduced by Tesnière in 1959 (Tesnière 1959). It is an important notion in dependency grammar. Valency is seen as the capacity a word has for combining particular patterns of other sentence constituents (Allerton 2005). More specifically, valency denotes the requirement that a word, usually a verb, noun, or adjective has on its complements, which can either be obligatory arguments or optional adjuncts (non-obligatory arguments) (Herbst 1988). So far, the existing research about valency of many languages is mainly concentrated on verbs, i.e. the ability of a verb to govern other words in a sentence (Tesnière 1959; Comrie 1993; Heringer 1993; Matthews 2007; Čech & Mačutek 2010). The verb opens up slots, in which the compliments enter as arguments (Heringer 1993: 303).

Huiyuan Jin, Zhengzhou University, Zhengzhou, P.R.China
Haitao Liu, Zhejiang University, Hangzhou; Guangdong University of Foreign Studies, Guangzhou, P.R.China, htliu@163.com

https://doi.org/10.1515/9783110573565-006

The concept of valency originated from chemistry. As the founder of modern valency, Tesnière intended valency to have dual properties: semantic and syntactic (Ágel 2000; Liu 2009). Semantic components determine the semantic roles (e.g. agent, patient) and the semantic class membership (e.g. ±human, ±animate) of participants. Semantic valency of verbs is primary, which results from their inherent giveness. The syntactic component determines the numbers and the syntactic forms of participants, which assures the correctness of syntactic positions.

However, observing only the semantic and syntactic aspects of verb valency is not sufficient. For example, the selection of a participant as the subject is not only dependent on the verb, but also influenced by the discourse structure. Though a complete sentence with all obligatory participants and actants is grammatically correct, sometimes, it is not necessarily appropriate in communications. The selection of the subject is mostly predictable. Therefore, language users tend to omit the word/form of the subject and the sentence/utterance is still understandable, especially in Chinese. This aspect of valency which indicates which participant of the verb is selected as part of the sentence form, is called the pragmatic valency. Pragmatic aspects of verb valency require a discourse or contextual observation. It is based on semantic-syntactic valency, but reflects the actual valency of a verb in real discourse context. Pragmatic analysis of verb valency belongs to the dynamic observation, which indicates that verb valency is also dependent on real communication circumstances or context.

Previous research into valency mostly concentrates on the quantitative investigations in a number of languages, e.g. English, Chinese, German, etc. Köhler (2005) made the first empirical attempt to observe the quantitative properties of German verb valency and reported some regular distributions of unique valency patterns, of complement variants (fitting the modified Zipf-Alekseev distribution), and of complementation patterns of each verb (fitting the Zipf-Mandelbrot distribution). Following Köhler's method, Čech et al. (2010) examined the distribution of complementation patterns in Czech, finding that the shorter the verb is, the more verb complementation patterns it has. Liu (2011) obtained similar results in English verb valency: the number of variants of English verbs follows the modified Zipf-Alekseev distribution; and the greater the valency of a verb is, the shorter the verb is, etc. Gao et al. found that the rank-frequency distributions of both valency polysemy of the 500 most frequent verbs in contemporary Chinese abide by a power law; and that valency and polysemy of these verbs abide by Good distribution and the positive negative binomial distribution respectively (Gao et al. 2014). However, these researches

on verb valency are mostly static quantitative descriptions. As a matter of fact, they are all based on the preliminary definition of valency. Researches into dynamic valency patterns are thus overlooked by most researchers.

Semantic and syntactic valencies of a verb are observed from a static perspective, whereas pragmatic aspects need to be observed from a dynamic perspective, i.e. from the real context. Ellipsis is a very frequent linguistic phenomenon in natural human languages, especially in spoken languages. The definition of ellipsis from a distinguished Chinese linguist, Lv Shuxiang, is provided here: "Ellipsis is confirmed only when the following two requirements are satisfied: first, a sentence, unless certain words are added to it, is ambiguous without context; second, the complementary words could have existed, and there is only one possibility for the complementary words. (Lv 1979: 67)" Unless both requirements are satisfied, the linguistic phenomenon cannot be identified as ellipsis.

Thus, elliptical constituents provide researchers with a good opportunity to observe the pragmatic valencies of verbs. According to the number of the obligatory valencies of a verb, verbs can be divided into four categories, i.e. zero-valency verbs, one-valency verbs, two-valency verbs, and three-valency verbs. According to their actual use in real context, the pragmatic valencies could also be marked off manually. Thus, if we build a corpus with the valency information annotated, we might then find out the elliptical patterns of verbs based on the notion of valency. Therefore, a corpus of modern spoken Chinese is established in current research, in which the frequently appearing elliptical phenomena are marked and analyzed according to valency. Based on the relevant information above, this article then attempts to explore the following three research questions:

1) Verbs can be divided into four categories according to the number of their obligatory valencies, and the current research will only take the verbs involved in the elliptical phenomena into consideration. Then what are the rank-distribution principles of these ellipsis-related verbs with different categories?
2) Verbs with different valencies might display different elliptical patterns. What are the rank-frequency distributions and elliptical patterns of these different categories of verbs?
3) The elliptical constituents of the verbs are generally subjects and objects. Therefore, which constituent is more frequently omitted in the oral discourse? And what are the distribution patterns of the elliptical constituents of the verbs?

This article will then introduce the materials and methods in the second section. The third section presents the results and relevant discussions. The last section is the conclusion part, together with relevant limitations and future research.

2 Research methods and materials

Spoken language, as the source and origin of human language, reflects the most active and real perspectives of natural languages (Jespersen 1924). Elliptical phenomena widely exist in numerous languages, especially in their spoken forms. For the considerations above, a modern spoken Chinese corpus was built for the current research. Since it is very difficult to select and record natural spoken language in people's daily life, natural dialogues in various programs on radio or on TV are used here. Dialogues between hosts/hostesses and distinguished guests in those programs were transcribed. There are roughly 10,000 Chinese words (roughly 20,000 Chinese characters) in the spoken Chinese corpus, which are randomly selected from the following programs, namely, *A Date with Lu Yu* , *Starface*, *Topic for Today*, *Current Affairs*, *Facing Joy*, *Auto World*, *Traffic Announcement*, *Legal Outlook on Society*, *People's Story*. More detailed information of the corpus is listed in Tab. 1.

Tab. 1: The sources of the corpus and their basic information

Order	Titles of Programs	Number of Texts	Number of Characters	Number of words
1	A Date with Lu Yu	9	25,387	13,874
2	Starface	6	16,387	8,749
3	Topic for Today	7	21,340	12,670
4	Current Affairs	6	20,871	11,642
5	Facing Joy	5	17,483	9,854
6	Auto World	6	20,378	11,092
7	Traffic Announcement	8	23,761	12,739
8	Legal Outlook on Society	7	22,487	12,873
9	People's Story	6	15,362	9,376
Total		60	183,456	102,869

Elliptical phenomena frequently appear in the spoken language. After the construction of the spoken Chinese corpus, we need to mark off the elliptical constituents based on the definition of ellipsis provided in Section 1, Introduction. As previously mentioned, verbs can be divided into four categories according to the number of their obligatory valencies, i.e. zero-valency verbs, one-valency verbs, two-valency verbs, and three-valency verbs. Zero-valency verbs indicate that they do not have any actants. In modern Chinese, these verbs are often related to weather or meteorology, such as "下雪[1] (*xia xue*, snow)", or "下雨 (*xia yu*, rain)". In terms of their quantity, they are far smaller than the other three categories. One-valency verbs refer to those that can only have one obligatory actant, with the actant being the subject in the sentence structure. One-valency verbs are usually intransitive verbs in Chinese, such as "休息 (*xiu xi*, rest)", "游泳 (*you yong*, swim)", etc. Similarly, two-valency verbs refer to those that can only have two obligatory actants, one being the subject, and the other, the object. They are usually transitive verbs in Chinese, such as "喜欢 (*xi huan*, like)", "唱 (*chang*, sing)", etc. Three-valency verbs refer to those that could have three obligatory actants, the first one being the subject, the second one being the direct object, and the third being the indirect object, such as "给 (*gei*, give)", or "帮助 (*bang zhu*, help)" in modern Chinese.

According to the definition of ellipsis, every elliptical case could be marked off, which is symbolized by the capitalized letter "R". In addition, the elliptical constituents might be continuous throughout the discourse. Thus we also marked their appearing sequence accordingly, right after the letter "R". Then a dependency syntax approach is used here to analyze the syntactic relationship between each word in a sentence. After the annotation of the dependency relations, the valencies of verbs also need to be annotated. We use Microsoft Excel to accomplish the annotation process, as shown in Tab. 2.

Tab. 2 displays the annotations of the dependency relationships between each word in a sentence and different kinds of verb valency. The second and forth columns are the word classes of Dependents and Governors; the fifth column shows the dependency relationships between the Dependents and the Governors. The sixth column shows the syntactic-semantic valency of the verbs of the Governors.

1 In the following bracket, the Chinese *pinyin* and the corresponding English translation are both provided. It is the same case with the following Chinese characters.

Tab. 2: The annotation of dependency relations and verb valency[2]

Dependent		Governor		Dependency relationships	Syntactic-semantic valency	Pragmatic valency	Verbs' omission
Text	Pos	Text	Pos				
戏迷-R1	n	唱	v	subj	2	0	s
您-R1	rr	说-R1	v	subj	2	1	s
岁	qt	说-R1	v	obj	2	1	s
您-R2	rr	说-R2	v	subj	2	1	s
岁	qt	说-R2	v	obj	2	1	s
您-R3	rr	说-R3	v	subj	2	1	s
岁	qt	说-R3	v	obj	2	1	s
他-R1	rr	出生	vi	subj	1	0	f
他-R1	rr	上	v	subj	2	1	f
他-R1	rr	是-R1	vshi	subj	2	0	s
我-R1	rr	见到	v	subj	2	1	f
我-R2	rr	见到-R1	v	subj	2	1	s

Note: Pos is short for part of speech, which means the word classes.

According to Section 1 Introduction, the syntactic-semantic valency refers to the static and maximum state of a verb valency. We looked it up in *the Semantic Dictionary of Chinese* that was compiled by the Center for Chinese Linguistics PKU[3]. The information provided by the semantic dictionary is displayed in Tab. 3, where we give examples of one-valency verbs, two-valency verbs, and three-valency verbs, respectively.

As shown in Tab. 3, this Chinese semantic dictionary provides the following information: the number of meanings, number of valencies, relevant information of the subject and the object, among which the third column, i.e. the number of valency, is what we need. The number of valency displays the syntactic-semantic valency of a verb, which describes verb valency from a static perspective. The dynamic description of a verb valency is defined by its pragmatic valency, which is annotated in the seventh column in Tab. 2. The last column in Tab. 2 indicates whether the verbs themselves are omitted together

2 Please see Appendix for the explanations of the abbreviation symbols in Tab. 2.

3 This semantic dictionary, which was compiled by Center for Chinese Linguistics in PKU, can be used online. The webpage is http://ccl.pku.edu.cn/ccl_sem_dict/.

with their actants. If yes, then an "s" was marked up; if no, then an "f" was marked up.

Tab. 3: The syntactic-semantic verb valency for one, two, and three-valency verbs

Verbs	Number of meanings	Number of valencies	Subject	Object	Object	Examples
出生(*chu sheng*)born	1	1	Person			这个孩子出生了。 The child was born.
说(*shuo*) speak	1	2	Person	Specific affairs/abstract process or affairs		他说英文。 He speaks Chinese.
给(*gei*) give	1	3	Person	Specific or abstract affairs	Person	我给他两本书。 I give him two books.

3 Results and discussions

3.1 The rank-frequency distribution of ellipsis-related verbs

This research attempts to study a series of elliptical patterns of verbs. As shown in Tab. 2, only the verbs involved in elliptical phenomenon are studied here. According to the statistical results, there are 1733 verbs in the elliptical phenomena in the spoken Chinese corpus. The number of each category and their percentages are listed in Tab. 4.

As shown in Tab. 4, the frequency of one-valency verbs is 150, two-valency verbs, 1,572, and three-valency verbs, 11. There are no zero-valency verbs in our corpus. The number of two-valency verbs occupies 90.71 percent of the whole verbs total number, which indicates that the number of two-valency verbs is the largest among the three categories. Previous studies hypothesize that the rank-frequency distribution of valency abides by a common rank frequency distribution or function (Köhler and Altmann 2009). These studies show that the relationship between rank-order and frequency of verbs carrying different valencies mostly displays a power-law trend, i.e.

$$y = ax^b \tag{1}$$

Tab. 4: Numbers of each verb category and their percentages

Categories of verbs	Frequency	Percentage
One-valency	150	8.66%
Two-valency	1572	90.71%
Three-valency	11	0.63%
Total	1733	100.00%

When verbs are omitted, then how about the rank-frequency distributions of verbs with different valencies? The frequency of elliptical verbs with one, two, and three valencies is listed in Tab. 5.

Tab. 5: Frequency of elliptical verbs with one, two, and three valencies

One-valency verbs		Two-valency verbs		Three-valency verbs	
Rank	Frequency	Rank	Frequency	Rank	Frequency
1	10	1	184	1	3
2	9	2	93	2	2
3	7	3	77	3	1
4	5	4	58	4	1
5	4	5	36	5	1
6	4	6	35	6	1
7~10	3	7	31	7	1
11~31	2	…	…		
32~88	1	162~335	1		

According to the statistical results displayed in Tab. 5, we then made the fitting test to observe whether their rank-frequency distributions obey the power law, too. Their fitting results are presented in Fig. 1, Fig. 2, and Fig. 3.

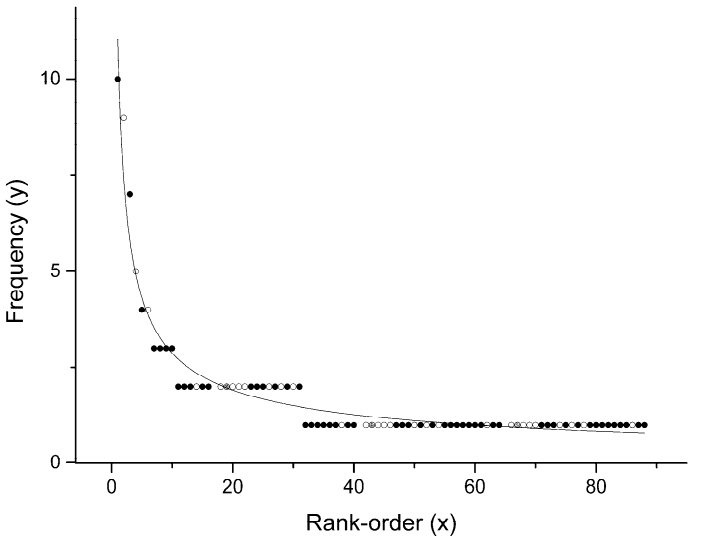


Fig. 1: Rank-frequency distribution of elliptical one-valency verbs

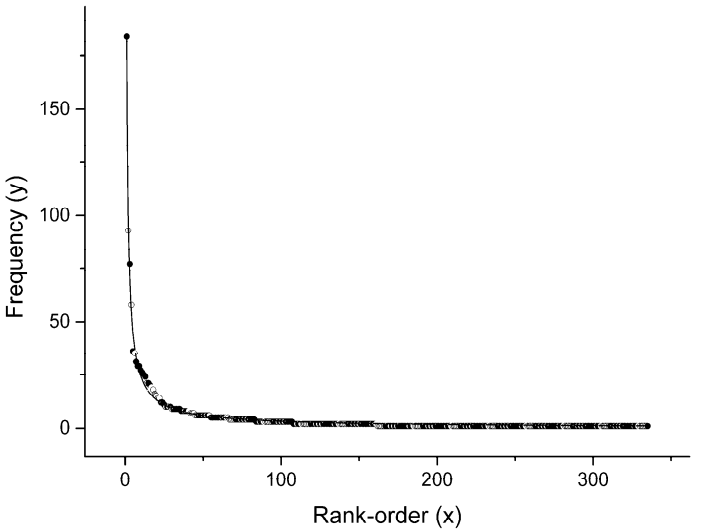


Fig. 2: Rank-frequency distribution of elliptical two-valency verbs

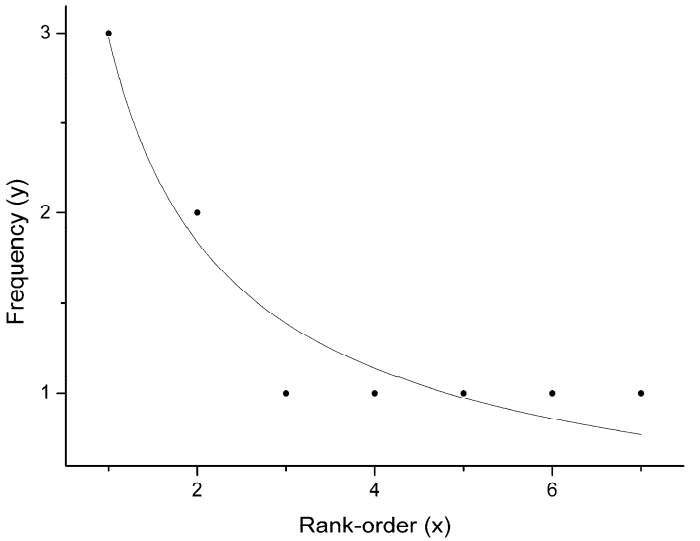


Fig. 3: Rank-frequency distribution of elliptical three-valency verbs

As shown above, the rank-frequency distribution of verbs with one, two, and three valencies all obey the power law, with their coefficient of determination R^2 being 0.94, 0.99, and 0.93, respectively. The values of the parameters a and b for Fig. 1 are 11.04 and −0.59, 183.41 and −0.88 in Fig. 2, and 2.98 and −0.96 in Fig. 3, respectively. These fitting results help to answer the first research question, i.e. the rank-frequency distribution of elliptical verbs obeys the power law. Similar to the fitting results of other research findings of the rank-frequency distribution of verb valencies, it seems that the power law is popular in the area of verb valency studies.

3.2 Distribution and the elliptical patterns of verbs with different valencies

As shown in Tab. 2, the syntactic-semantic valencies and the pragmatic valencies of verbs are marked off in the annotated corpus. Based on different numbers of syntactic-semantic valencies, different numbers of pragmatic valencies, and the omission of verbs *per se*, we then summarized 9 kinds of possible combinations, and calculated their frequencies and corresponding percentages in each case, as presented in Tab. 6.

Tab. 6: Possible combinations of elliptical patterns of three categories of verbs

Possible combinations	Syntactic-semantic valency	Pragmatic valency	Omission of verbs	Frequency	Percentage
1	2	1	f	993	63.17%
2	2	0	f	257	16.35%
3	2	0	s	183	11.64%
4	2	1	s	139	8.84%
5	1	0	f	127	84.67%
6	1	0	s	23	15.33%
7	3	2	f	6	54.55%
8	3	1	f	4	36.36%
9	3	1	s	1	9.09%

According to the statistical results displayed in Tab. 6, we then made the fitting test and we found that their rank-frequency distribution obeys the power law, with the parameter a=987.30, b=−1.70, and R^2=0.99, as shown in Fig. 4:

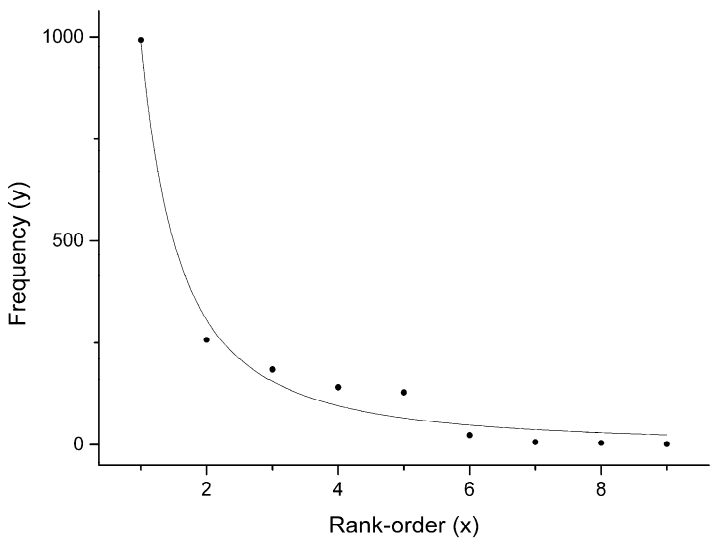


Fig. 4: Rank-frequency distribution of the elliptical patterns of the three categories of verbs

We will then describe and analyze each category of verbs as follows. For one-valency verbs, there is only one possibility of their actants' omission, i.e. the omission of their only actant, which functions as the subject in the sentence structure. However, there are still different cases when considering the omission of verbs *per se*. As shown in Tab. 6, we can find that, for most one-valency verbs, their elliptical actants would be omitted without verbs themselves. This case occupies 84.67% of the total number. However, for a small number of one-valency verbs, their elliptical actants would be omitted with verbs themselves. The percentage of this case is about 15%.

For two-valency verbs, most of them are more inclined to omit one of its obligatory valency constituents, and the two-valency verbs themselves tend to be reserved together with the remainder. This elliptical pattern can be described as the "21f"[4] model, as displayed in the *x* axis in Fig. 4. The frequency of the "21f" elliptical model is 993 in our corpus. The frequency of the "20f" elliptical pattern, indicating that both obligatory valencies are omitted, ranks the second among all the two-valency verbs. The frequency of "20s" elliptical pattern is 183, implying that both obligatory valency constituents and the two-valency verbs themselves are omitted. The last case is "21s", indicating that one of the obligatory valency constituents and the two-valency verbs themselves are omitted during communication, whereas the other obligatory valency constituents are reserved. The frequency of this pattern is 139, ranking the last among the four cases. It should also be noted that, the frequency of the elliptical pattern of "21f" in two-valency verbs is the highest among the three categories of verbs, implying that this pattern is probably the most frequently used elliptical model in modern spoken Chinese.

For three-valency verbs, it is very obvious that the trend of ups and downs of three-valency verbs is not as abrupt as the other two categories. As the statistical results show, there are three kinds of combinations, namely "32f", "31f", "31s", in which "32f" is the most frequently used pattern. The frequency of "32f" pattern is 6, and "31f" is 4, while "31s" is 1. Obviously, the total number of three-valency verbs is much fewer than the other two categories. Similarly, three-valency verbs are also inclined to omit one of its constituents of obligatory va-

4 The "21f" model implies the possible combinations, which includes the information in the second, the third, and the fourth columns in Tab. 6. "2" means the number of the syntactic-semantic valency; "1" refers to the number of pragmatic valency; and "f" means the verb is omitted in the context. It is the same case with the following content in this article, such as "32f", "31f", etc.

lency in communication. In addition, three-valency verbs also tend to omit two obligatory valency constituents and reserve the verbs themselves.

The description and analysis of the elliptical patterns for verbs with different valencies yield the answers to the second research question. It could be concluded that, on the one hand, the distribution of the elliptical patterns for the three categories of verbs obey the power law. On the other hand, the three categories of verbs display similar elliptical patterns to some degree, yet discrepancies still exist among them. It seems that all of them are more inclined to omit one of their obligatory constituents, in which the "21f" model of two-valency verbs is the most frequently used elliptical pattern in oral context. The elliptical principle of "omitting the old information and reserving the new information" might be the potential reason for these elliptical patterns.

3.3 The distribution pattern of the dependency relationship of the elliptical valencies

In this section, we attempt to find the possible answer to the third question. In real context, obligatory valencies are not necessarily satisfied, especially when they are repetitive with the constituents that have appeared in preceding sentences or clauses. The statistical results in the previous section probably indicate that three categories of verbs are all more inclined to omit one of their obligatory semantic valencies in dynamic communication process. Semantically speaking, complete valencies for verbs usually contain agents and patients, whereas for three-valency verbs, they also contain datives. Syntactically speaking, complete valencies for verbs generally include subjects and objects, whereas for three-valency verbs, they also contain object 2 (e.g. indirect object). According to the syntactic dependency treebank that we built, we would be more interested in the following question: which constituent, namely, whether the agent or the patient, or whether the subject or the object, is more frequently omitted in the communication discourse for different categories of verbs? This question is actually in consistent with the third question in Section 1 Introduction.

We calculated the frequency of different syntactic constituents for different categories of verbs, as shown in Tab. 7.

Tab. 7: The frequency of dependency relations of the elliptical constituents in the three categories of verbs

Syntactic constituents	Corresponding implications	Fre.	%
subj	subject	1272	73.40%
obj	object	376	21.70%
obja	object of volitive auxillary	34	1.96%
sentobj	object of Chinese clause	33	1.90%
subobj	object of pivotal construction	8	0.46%
obj1	direct object	5	0.29%
obj2	indirect object	2	0.12%
soc	complement of pivotal construction	2	0.12%
baobj	object of Chinese character "把（*ba*)"	1	0.06%

Fig. 5 shows that there is an excellent fit of the power law to the frequency of the dependency relations of the elliptical constituents in the three categories of verbs, with the determination coefficient R^2 being 0.9853, the parameter a being 1280.96, and the parameter b, −2.17.

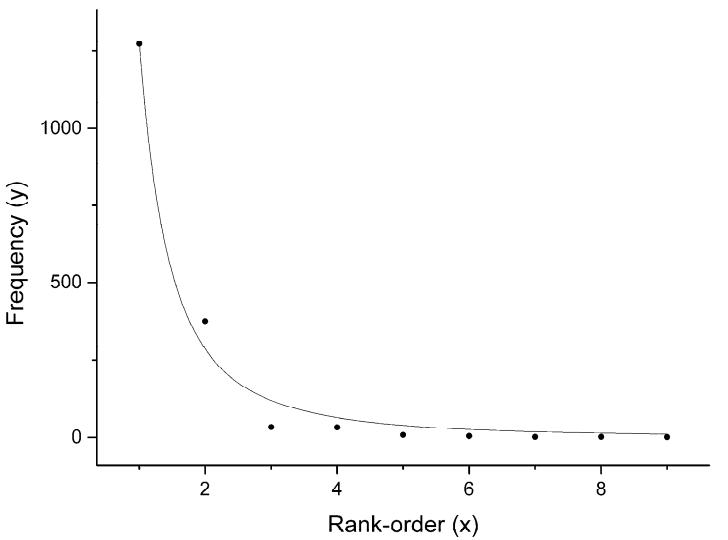


Fig. 5: Fitting the power-law to the data in Tab. 7

As shown above, apparently, if one of its obligatory valencies is omitted, chances are most of them will be the subjects in terms of syntax or the agents semantically. The frequency of the omitted subjects is 1,272, ranking the highest among all the elliptical constituents. The second highest elliptical constituent is, of course, the objects, whose frequency is 376. In addition, the dependency relations of "obja", "sentobj", "subobj", "obj2", "obj1", "soc", "baobj" are actually different manifestations of verbs' objects in different sentences with different syntactic structures. However, Tab. 7 and Fig. 5 provide us with a general observation of the elliptical dependency relations for all the three categories of verbs. Then it is necessary to observe the elliptical phenomenon for each category, as shown in Tab. 8.

Tab. 8: Frequency of elliptical valencies for the three categories of verbs

Elliptical valencies	Corresponding implications	1-valency verbs		2-valency verbs		2-valency verbs	
		Fre.	%	Fre.	%	Fre.	%
subj	subject	150	100	1,123	60.31	3	27.27
obj	object			373	20.03	1	9.09
obja	object of volitive auxillary			322	17.29		
sentobj	object of Chinese clause			32	1.72	1	9.09
subobj	object of pivotal construction			7	0.38	1	9.09
obj1	direct object			2	0.11		
soc	complement of pivotal construction			2	0.11		
baobj	object of Chinese character "把（*ba*)"			1	0.05		
obj2	indirect object					5	45.45

In Tab. 8, subjects are the only elliptical valencies for one-valency verbs, whereas for two- and three-valency verbs, their elliptical constituents are plenty. The number of elliptical valencies for two-valency verbs is the highest among the three categories, whereas the number of those for three-valency verbs is too little to be statistically tested. Thus we also made a fitting test for the frequency of elliptical valencies for two-valency verbs. As shown in Tab. 8 and Fig. 6, there is also an excellent fit of the power law to the frequency of elliptical valency for two-valency verbs, with the determination coefficient R^2 being 0.96, the parameter a being 1133.17, and the parameter b, −1.67.

As shown in Tab. 8 and Fig. 6, both the frequency of elliptical valency for two-valency verbs and the frequency of the dependency relations of the elliptical constituents in the three categories of verbs perfectly fit the power law. For two-valency verbs, it is subjects, or agents, that are the most frequently elliptical constituents. Thus, compared with objects/patients, subjects are more likely to be omitted during communication. However, it is not the same case with three-valency verbs. Fig. 4 implicated that obj2 is the most frequently omitted constituent among the three obligatory valencies; whereas the number of elliptical subjects only ranks the second, far less than that of obj2. It seems that for three-valency verbs, subjects/agents are not their most frequently elliptical constituents during communication. This statistical result may not be consistent with our expectations, though, we should also be aware that the total number of three-valency verbs in our corpus is only 11, which is far too less to yield a very reliable result. Therefore, for a more general elliptical pattern of three-valency verbs, a larger number of collections of three-valency verbs are needed in future research.

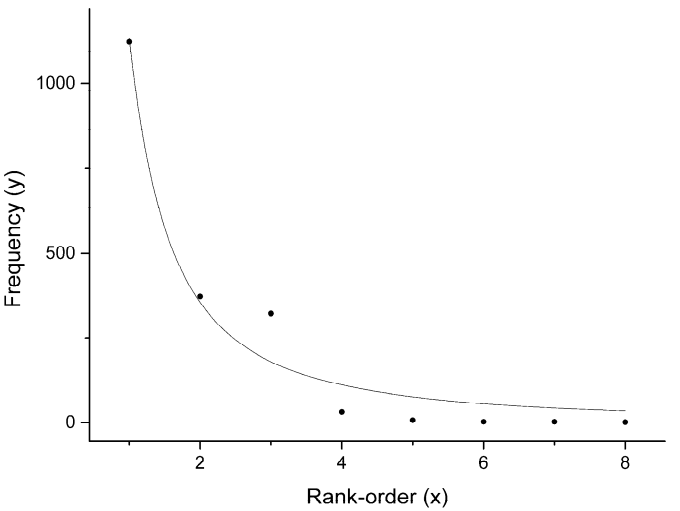


Fig. 6: Fitting the power law to the frequency of elliptical valency for two-valency verbs

4 Conclusions

The current research attempts to observe and describe the elliptical patterns of verbs from the notion of valency. Three main conclusions are obtained as follows.

First, we observed and quantitatively described the rank-frequency distribution of elliptical verbs with different valencies. Their statistical results showed that, their distributions obey the power law, with their coefficient of determination R^2 being 0.94, 0.99, and 0.93, respectively.

Second, we quantitatively described the elliptical patterns of the three categories of verbs, and the results also showed that they obeyed the power law. To be more specific, the three categories of verbs displayed similar elliptical patterns to some degree, yet discrepancies still exist among them. It seems that they are more inclined to omit one of their obligatory constituents, in which the "21f" model of two-valency verbs is probably the most frequently used elliptical pattern in oral context. The elliptical principle of "omitting the old information and reserving the new information" might be the potential reason for the observations.

Third, we statistically described the dependency relations of the elliptical valencies of verbs. Similarly, both the frequency of elliptical valencies for two-valency verbs and the frequency of the dependency relations of the elliptical constituents in the three categories of verbs fit the power law. In addition, the three categories of verbs tend to omit one of their obligatory valencies. For one-valency and two-valency verbs, the obligatory constituents are more likely to be subjects syntactically, or agents from semantic aspect; for three-valency verbs, the most frequently omitted constituents are usually the indirect objects (e.g. obj2). Since the number of three-valency verbs is inadequate in our research, a larger number of three-valency verbs and a larger corpus might contribute to a more reliable result in future research.

The current research was conducted based on a corpus of modern spoken Chinese. Though ellipsis is a very common linguistic phenomenon in various languages, discrepancies among different languages still exist. For example, subject ellipsis in English is not as common as it is in Chinese or Japanese. Verbs in different languages may display different elliptical patterns. Therefore, relevant studies of verb valencies from a dynamic perspective in more languages and relevant contrasts among these languages could be studied in future research.

References

Ágel, Vilmos. 2000. *Valenztheorie*. Tübingen: Narr.
Allerton, David. 2005. Valency grammar. In Keith Brown (Ed.), *The Encyclopedia of Language and Linguistics* (pp. 48–78). Kidlington: Elsevier Science Ltd.
Čech, Radek & Jan Mačutek. 2010. On the quantitative analysis of verb valency in Czech. In: P. Grzybek, E. Kelih & J. Mačutek (Eds), Text and Language. Structure, Functions, Interrelations (pp. 21–29). Wien: Preasen Verlag.
Čech, Radek, Petr Pajas & Jan Mačutek. 2010. Full valency, verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics, 17* (4), 291–302.
Comrie, Bernard. 1993. Argument structure. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (Eds.), *Syntax: Ein internationales Handbuch zeitgenössischer Forschung. Halbband 1* (pp. 905–914). Berlin: de Gruyter.
Gao, Song, Hongxin Zhang & Haitao Liu. 2014. Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics, 21* (1), 1–21.
Herbst, Thomas. 1988. A valency model for nouns in English. *Journal of Linguistics, 24* (2), 265–301.
Heringer, Hans. 1993. Basic ideas and the classical model. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld & Theo Vennemann (Eds.), *Syntax: Ein internationals Handbuch zeitgenössischer Forschung. Halbband 1* (pp. 298–316). Berlin: de Gruyter.
Jespersen, Otto. 1924. *The Philosophy of Grammar*. London: George Allen & Unwin Ltd.
Köhler, Reinhard. 2005. Quantitative Untersuchungen zur Valenz deutscher verbena. *Glottometrics*, *9*, 13–20.
Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.
Liu, Haitao. 2011. Quantitative properties of English verb valency. *Journal of Quantitative Linguistics, 18* (3), 207–233.
Lv, Shuxiang. 1979. *The Problems of Chinese Grammar Analysis*. Beijing: The Commercial Press.
Matthews, Peter. 2007. The scope of valency in grammar. In Thomas Herbst & Katrin Gotz-Votteler (Eds.), *Valency: Theoretical, Descriptive and Cognitive Issues* (pp. 3–14). Berlin: de Gruyter.
Tesnière, Lucien. 1959. *Elements de Syntaxe Structural*. Paris: Klincksieck.

Appendix

Tab: The corresponding implications of the abbreviations in Tab. 2

Abbreviations	Corresponding implications	Abbreviations	Corresponding implications
n	noun	subj	subject
rr	pronoun	obj	object
qt	quantifier	s	yes
v	verb	f	no
vshi	be		

Haruko Sanada

Negentropy of Dependency Types and Parts of Speech in the Clause

Abstract: As one of our studies of the quantitative aspects of Japanese valency, in this paper we focus on the amount of information in the clause. The possibility of choosing a word in the clause generally decreases as the position of the word moves from left to right. We studied the negentropy of dependency types (subject, object, time, place, predicate, and others) and of parts of speech in Japanese clauses. The results were compared to Köhler's study (2012). The negentropy of the parts of speech decreased more rapidly than that of the dependency types. We also found that there are combinations of dependency types with relatively strong connections such as time and place or a direct object and a predicate, and concluded that the occurrence of some dependency types does not depend on the absolute position in the clause, but on the relative position to other dependency types.

Keywords: valency; length; frequency; position; Japanese; Synergetic Linguistics

1 Aim of the Present Paper and Our Earlier Studies

This paper is one of our studies on the quantitative aspects of Japanese valency. We investigated (1) distributions of dependency types, (2) co-occurrence and the order of complements, (3) the Menzerath-Altmann law (relationships of lengths of two levels) among the sentence length, the clause length, the argument length (hereafter the term argument includes a complement, an adjunct and a predicate) referring to Köhler's study (1999), (4) relationships among length, position and depth of the clause, (5) relationships of the position of clauses in the sentence and length or number of arguments, and (6) linguistic properties to shorten the clause (Sanada 2012, 2014, 2015, 2016a, 2018a, 2018b, to appear).

Haruko Sanada, Rissho University, Tokyo, Japan, hsanada@ris.ac.jp

https://doi.org/10.1515/9783110573565-007

As part of a series on the quantitative characteristics of Japanese valency, in the present paper we focus on the amount of information carried from the beginning of the clause to the end. The probability of a word occurring in a certain clause generally decreases the further the position is to the end. It corresponds to Köhler's register hypothesis which is explained as "at each position of a given sentence, from left to right, the number of possible alternatives was determined". This was done first with respect to structural alternatives, then with respect to functional alternatives" (Köhler 2012: 85–86). At the beginning of the clause there are many choices of words or dependency types, but the possible alternatives become fewer if the position of a word moves to the end of a clause. For example, in English, the probability that a subject follows a predicate is much less than the probability that a predicate follows a subject. We investigated the negentropy of dependency types (subject, object, time, place, predicate and others) and of parts of speech (noun group, adjective-adverb group, verb group, interjection and conjunction group and others including postpositions, auxiliary verbs, prefixes and suffixes) in Japanese clauses when the position of a word moves from left to right.

In our last study (Sanada, to appear) we hypothesized that averages of the length of arguments in morphemes by dependency type in the clauses as y and the position of arguments in the clause as x have a significant relationship. The data showed that the average number of morphemes per argument decreases if the argument position moves to the end of the clause (Fig. 1). We also investigated the relationship between the position of the argument and the length of the arguments in morphemes, which are categorized by dependency type, i.e. the subject, the object, the place, the time, and the occasion (Fig. 2, Sanada, to appear). The data points shown with white markers in the figures are less than 10. However, none of the dependency types have a significant relationship with the argument position in the clause. Therefore, we concluded that the length of the argument with the dependency type which is closer to the end of the clause does not contribute to shortening the clause, even though the average length of all arguments decreases (Sanada, to appear).

In Fig. 1 and Fig. 2 averages of the argument lengths in morphemes were employed. Here the question arises how the distributions of the original data which was employed to obtain averages in Fig. 1 and Fig. 2 were obtained. The distributions of the data for each position in the clause must show the probability of dependency types (i.e. how many kinds of dependency types can be taken as an alternative) and their frequency which depends on the position in the clause. An interpretation can be made as to how information is shown in a clause if the position of a word moves from left to right: the meaning of a clause

is clearer and the rest of the information, i.e. the information that is not shown, is less if the position moves from left to right.

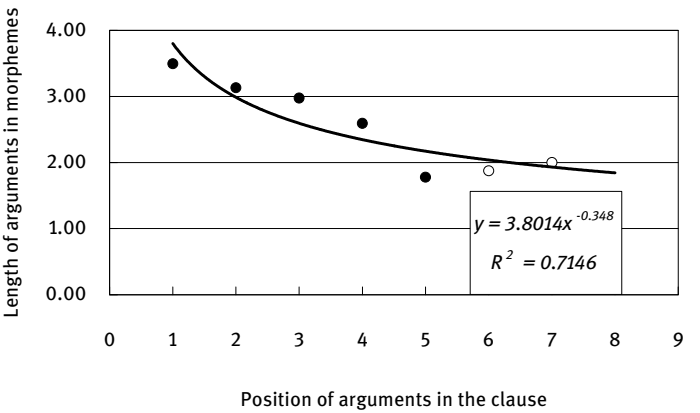


Fig. 1: Relationship between averages of the length of arguments in morphemes (*y*) and the position of arguments in the clause (*x*) for 243 "meet" clauses (*x*=1 as the beginning of the sentence)

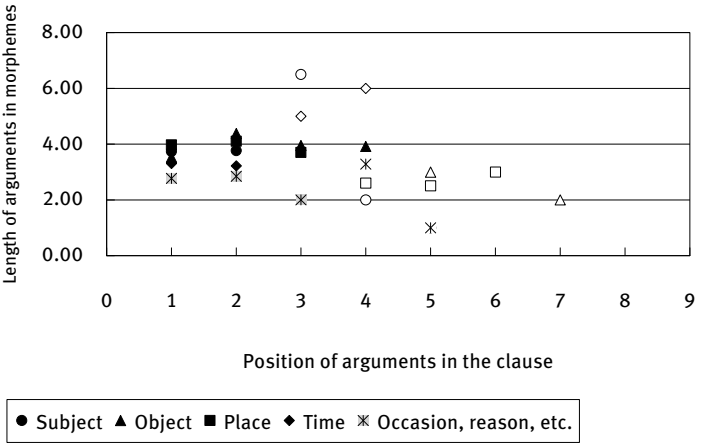


Fig. 2: Relationship between averages of the length of arguments in morphemes (*y*) by dependency type and the position of arguments in the clause (*x*) for 243 "meet" clauses (*x*=1 as the beginning of the sentence) (= Fig. 11c in Sanada, to appear)

In order to measure the decrease in information in the clause, we followed Köhler (2012: 84–92), and calculated the negentropy of the probability, i.e. a relative frequency, of possible arguments for the dependency type. The negentropy is defined as the entropy which takes a negative value. We also calculated the negentropy of the amount and probability of possible arguments for the part of speech in order to observe a decrease in choices of the part of speech in dependence with the position in the clause. Köhler (2012) investigated the negentropy of the number of alternatively possible constituent types (*ibid.* 87) and the negentropy of the probability of alternatively possible constituents (*ibid.* 91). However, we did not employ the negentropy of the amount of dependency types or that of the parts of speech, but rather the negentropy of the probability because individual dependency type, e.g. the object or the time, or individual part of speech, e.g. the noun or the adverb, have a different probability and do not have an equal chance to be taken. This point is also mentioned in Köhler (2012: 87).

Concerning the dependency type, Čech and Uhlířová (2014) listed 13 categories including the time and the place. We referred to their categories, and defined six categories such as the subject, the object, the predicate, the time, the place, and others (the occasion, the reason, the manner of the meeting, etc.), which were also employed in our former studies (Sanada 2016a, 2018a, 2018b, to appear). Frequency distributions of the dependency types of our data are shown in our studies (Sanada 2016a, to appear).

2 Definitions and Descriptions of Data

We employed the Japanese valency database (Ogino et al. 2003), which is the same as the one employed in our former studies (Sanada 2012, 2014, 2015, 2016a, 2018a, 2018b, to appear) because the present study is part of our series of valency studies. For the present study, 240 sentences were extracted from the valency database, including 243 clauses containing the verb “meet”. Three of the 240 sentences have two predicates with the verb “meet”. In the series of our valency study we originally focused on five verbs, i.e. “meet”, “tear”, “work”, “be born”, and “move”, in order to investigate grammatically various types of the verb which takes one, two or three valencies, etc. (See Sanada 2012), and results with the other verbs will be investigated.

The morphological analyzer *MeCab* (Graduate Schools of Informatics in Kyoto University; NTT Communication Science Laboratories 2008) and the digital dictionary *UniDic* (National Institute for Japanese Language and Linguistics

2008) were used to tag the parts of speech in our data, i.e. 2,365 morphemes with 765 arguments and 243 clauses which include the verb "meet". The software shows the boundary of the "short unit" (see below) as a morpheme. Errors were corrected by hand. Our former study took four linguistic levels, i.e. sentence, clause, argument and morpheme. The clause must have one predicate for each, and consist of complements, adjuncts, and a predicate. The present study follows these definitions.

The position of the argument in the clause is counted from the beginning of the clause. If the clause is divided into two parts by an embedded clause, arguments of the embedded clause are skipped and not counted.

The rules used for counting the data are shown here with an example, also shown in our former study. A space in the example shows a morpheme boundary. A single slash mark (/) and a double slash mark (//) show the boundary of arguments and boundaries of the clauses, respectively. The numbers of clauses, arguments and morphemes of the example follow their English equivalents. The numbers shown with "ID" after the example indicate a sentence number in the database.

(1) Abbreviations and grammatical remarks in the example

Abbreviations or grammatical remarks used in the example are as follows:

ATTR = attributive, GEN = genitive, OBJ = object, PAST = past tense form, SUBJ = subject.

(2) Definitions of the sentence and the morpheme

The "sentence" in Japanese is optically clear as it has a sentence marker at the end. The definition of "morpheme" is a topic that is still being discussed. We employ the definition of the "short unit" as a morpheme, which was developed by the National Institute of Japanese Language and Linguistics (National Language Research Institute 1964).

(3) Definition of the clause

The definition of "clause" is also a topic that is still being discussed in Japanese linguistics. Minami (1974, 1993) analyzed grammatically important types of Japanese clauses. Here, referring to his model, we studied our data using quantitative and empirical analyses. In the present study we defined it as a linguistic unit that has a predicate on the surface of the sentence.

(4) Definition of the argument

The argument is defined as the level between the clause and the morpheme. We regard the predicate and grammatical elements that are linked to the predicate as arguments in the clause. All elements in the clause are employed as arguments. Attributive elements, i.e. a noun and a postposition were treated as a part of the argument (see underlined words in Example 1). Among the argu-

ments, those tagged in the Japanese valency database (Ogino et al. 2003) were defined as the complement, the rest of the arguments, except the predicate, are defined as the adjunct. Conjunctions, except postpositions, which belong to the clause were also regarded as a member of the clause in the present study, and categorized as an adjunct. This definition should be discussed in our future studies.

Example 1:

(1) *Yogo no Ishikawa Keiko sensei wa*, /
nurse-teacher-ATTR Mrs. Keio Ishikawa-SUBJ, /
chugaku 3 nen no shojo no hahaoya ni/ *at ta*.
junior high school 3rd grade-ATTR girl-GEN mother-OBJ/ meet-PAST
"Mrs. Ishikawa Keiko of a nurse-teacher / met / a mother of the girl in the 3rd grade of the junior high school."
Clause=1, Argument = 3, Morpheme=16.

For more detailed definitions, e.g. definitions related to the predicate or other special cases, Sanada (2016a) should be referred to.

3 Distributions of Number of Arguments by Dependency Type and Their Negentropy

Employing 765 arguments with 243 clauses which include the verb "meet", we can confirm that the distribution of the number of arguments is a function of the position in the clause (Fig. 3, Tab. 1). The number of arguments in Tab. 1 corresponds with data points which were employed to obtain the averages of the argument lengths in Fig. 1. For example, the average of the argument length is 3.49 for the first argument in the clause (the position (x) = 1) in Fig. 1, and the total number of data for arguments in the first position of the clause is 243 which is shown in Fig. 3 and Tab. 1.

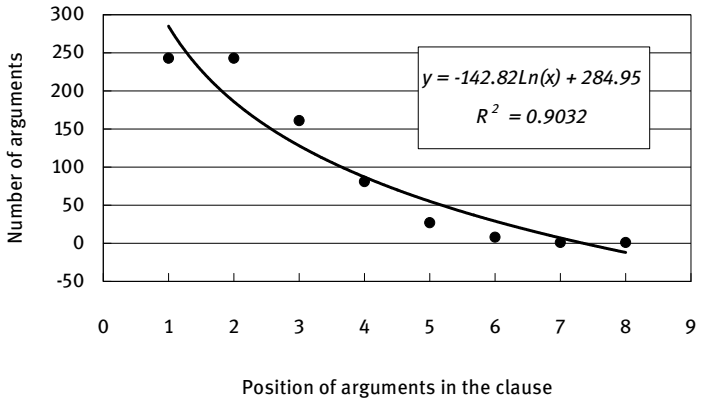


Fig. 3: Distribution of the number of arguments as a function of the position in the clause

Tab. 1: Distributions of the number of arguments by position in the clause

Position of argument in the clause (1 as the beginning of the clause)	Number of arguments
1	243
2	243
3	161
4	81
5	27
6	8
7	1
8	1
Total	765

We investigated the distributions of the number of arguments by dependency type and by position (Fig. 4 and Tab. 2). The total amount of arguments of all dependency types for each position corresponds to the amount of arguments in Tab. 1. We followed Köhler (2012: 84–92), and sorted the numbers of arguments in Tab. 2 by descending frequency, and calculated relative frequencies (p) for the positions (Tab. 3). For example, in Tab. 2 the frequency of the Object is 70 in the first position of the clause, and it is the most frequent dependency type in the first position. There are a total of 243 arguments for the first position, therefore the relative frequency of the Object is 0.2881 (= 70 / 243), which is shown in

Tab. 3. We can observe in Tab. 3 that relative frequencies of dependency types for each position are not homogeneous, and that the probabilities of the possible alternatives of dependency types are not equal. Therefore, we calculated the negentropy of the probability of dependency types, as mentioned above.

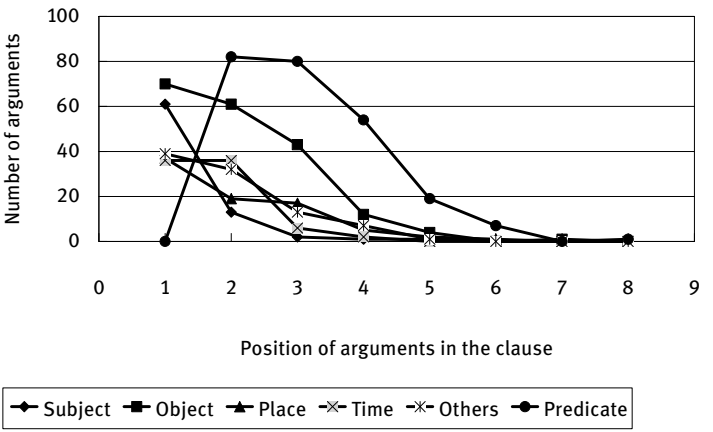


Fig. 4: Number of arguments by type and by argument position in the clause

Tab. 2: Number of arguments by dependency type and by the argument position in the clause

Argument position in the clause	Number of Subject	Number of Object	Number of Place	Number of Time	Number of Others	Number of Predicate	Total
1	61	70	37	36	39	0	243
2	13	61	19	36	32	82	243
3	2	43	17	6	13	80	161
4	1	12	5	2	7	54	81
5	1	4	2	0	1	19	27
6	0	0	1	0	0	7	8
7	0	1	0	0	0	0	1
8	0	0	0	0	0	1	1
Total	78	191	81	80	92	243	765

We also calculated the negentropy ($-H$) employing the following formula (*ibid.* 91):

$$H = -\sum P_i \ln P_i \quad (1)$$

with the relative frequency p_i and the logarithm to base e. Fig. 5 shows the relationship between the negentropy and the position of the argument in the clause. The value of the negentropy for the second position is the highest among the values of negentropy for eight positions, and it is followed by the negentropy value for the first position. The values of negentropy rapidly decrease for the third or closer positions to the end of the clause. It can be interpreted that the first or second position of the clause has the most possible alternatives for the dependency types.

Tab. 3: Number of arguments and their relative frequencies by position sorted by descending frequency

Pos. 1			**Pos. 2**			**Pos. 3**		
Type	*n*	*p*	Type	*n*	*p*	Type	*n*	*p*
Object	70	0.2881	Predicate	82	0.3374	Predicate	80	0.4969
Subject	61	0.2510	Object	61	0.2510	Object	43	0.2671
Others	39	0.1605	Time	36	0.1481	Place	17	0.1056
Place	37	0.1523	Others	32	0.1317	Others	13	0.0807
Time	36	0.1481	Place	19	0.0782	Time	6	0.0373
			Subject	13	0.0535	Subject	2	0.0124
Sigma(*n*)	243		Sigma(*n*)	243		Sigma(*n*)	161	
−Sigma(*H*)		1.5686	−Sigma(*H*)		1.6193	−Sigma(*H*)		1.3178

Pos. 4			**Pos. 5**			**Pos. 6**		
Type	*n*	*p*	Type	*n*	*p*	Type	*n*	*p*
Predicate	54	0.6667	Predicate	19	0.7037	Predicate	7	0.8750
Object	12	0.1481	Object	4	0.1481	Place	1	0.1250
Others	7	0.0864	Place	2	0.0741			
Place	5	0.0617	Subject	1	0.0370			
Time	2	0.0247	Others	1	0.0370			
Subject	1	0.0123						
Sigma(*n*)	81		Sigma(*n*)	27		Sigma(*n*)	8	
−Sigma(*H*)		1.0824	−Sigma(*H*)		0.9671	−Sigma(*H*)		0.3768

Pos. 7			**Pos. 8**		
Type	*n*	*p*	Type	*n*	*p*
Object	1	1.0000	Predicate	1	1.0000
Sigma(*n*)	1		Sigma(*n*)	1	
–Sigma(*H*)		0.0000	–Sigma(*H*)		0.0000

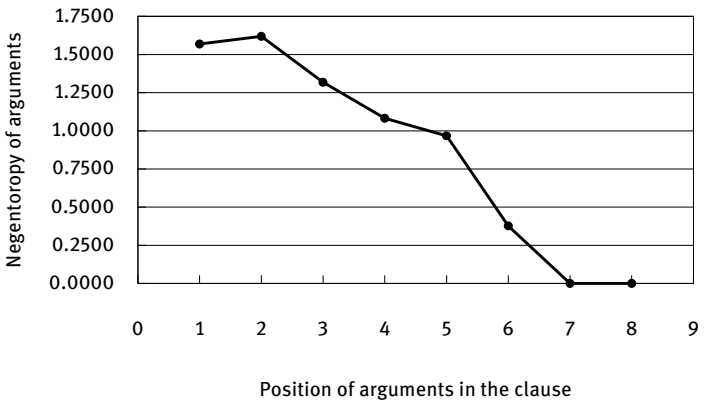


Fig. 5: Negentropy of arguments and the position of arguments in the clause

4 Order of Parts of Speech in the Argument and Their Negentropy

As mentioned above, we employed the morphological analyzer *MeCab* (Graduate Schools of Informatics in Kyoto University; NTT Communication Science Laboratories 2008) and the digital dictionary *UniDic* (National Institute for Japanese Language and Linguistics 2008). The digital dictionary *UniDic* (National Institute for Japanese Language and Linguistics 2008) defined over 40 subgroups of the parts of speech in total, including some symbols (Sanada 2016c, to appear). Referring to Kabashima's studies (1954, 1955, 1957), we summarized the subgroups into the following five groups:

- N: nouns and pre-nouns,
- A: adverbs, adjectives and adjective verbs,

- V: verbs,
- I: interjections and conjunctions, and
- O: others including postpositions, auxiliary verbs, prefixes and suffixes.

Tab. 4: List of groups and parts of speech employed by the morphological analyzer *UniDic*

Group	Part of Speech	Part of Speech
N	Pre-noun	noun: proper noun: place name: country
	noun: proper noun: general	noun: auxiliary verb type
	noun: proper noun: personal name: general	noun: numeral
	noun: proper noun: personal name: family name	noun: general: Suru ending type
	noun: proper noun: personal name: first name	noun: general: Suru ending and adjective verb type
	noun: proper noun: organization	noun: general: general
	noun: proper noun: place name: general	noun: general: adjective verb type
		noun: general: adverb type
V	verb: general	verb: auxiliary verb type
A	adjective verb: Tari ending	adjective: general
	adjective verb: general	adjective: auxiliary verb type
	adjective verb: auxiliary verb type	adverb
		pre-noun adjectival
I	interjection: filler	conjunction
	interjection: general	
O	postposition: case	suffix: adjective type
	postposition: linking	suffix: verb type
	postposition: ending	suffix: noun type: Suru ending
	postposition: attributive	suffix: noun type: general
	postposition: conjunctive	suffix: noun type: adjective verb type
	postposition: adverbial	suffix: noun type: counter
	auxiliary verb	suffix: noun type: adverb type
	prefix	symbol: script
	suffix: adjective verb type	symbol: general

The list of our five groups and the parts of speech employed by the morphological analyzer *UniDic* are shown in Tab. 4. Kabashima (1954, 1955, 1957) studied relationships between genres of text and quantitative aspects of four groups of parts of speech, i.e. Noun group, Adverb group, Verb group, and Interjection or conjunctions group. His work is responsible for a law in quantitative linguistics known as Kabashima's Function (Sanada 2016b). We found Kabashima's definition of the four groups of parts of speech to be useful if a quantitative analysis of morphologies is to be carried out. This is due to the fact that he defined the

groups of parts of speech while considering quantitative characteristics of morphemes, characteristics of text, and grammatical characteristics of morphemes. The group O, described above, is not included in Kabashima's definitions, and it is mainly for functional words. For our study, we had to carefully consider whether the group O should be included in the following sections or not, as its linguistic characteristics are different from the other four groups.

We confirmed distributions of the amount of morphemes including group O and excluding group O in dependence with the position of the morpheme in the argument, shown in Fig. 6. The distributions of the parts of speech in dependence with the position of morphemes in the argument are also shown in Fig. 7. Detailed numbers of the parts of speech in dependence with the position of morphemes in the argument are shown in Tab. 5.

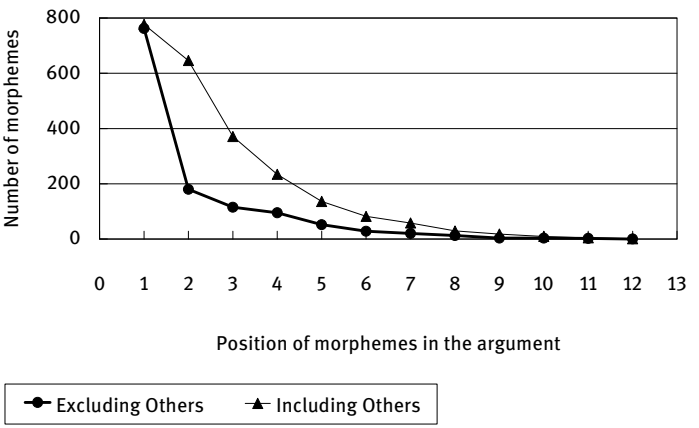


Fig. 6: Distribution of the number of morphemes (*y*) by the position of the morpheme in the argument (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the argument)

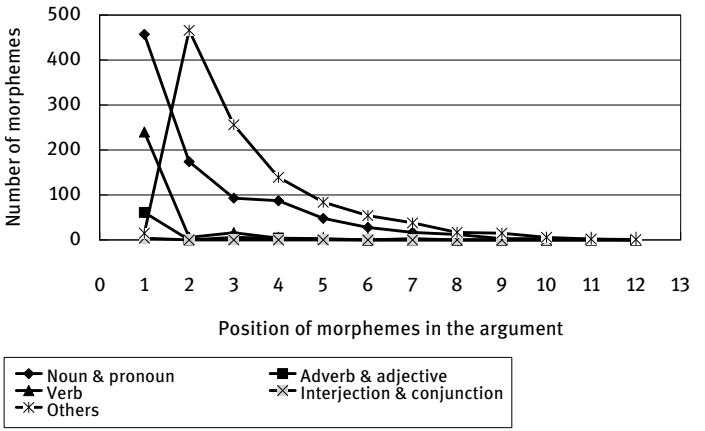


Fig. 7: Distributions of numbers of parts of speech (*y*) for each position of morphemes in the argument (*x*) with 243 "meet" clauses (*x*=1 as the beginning of the argument)

Tab. 5: Numbers of morphemes for groups of the part of speech and for the position in the argument

Position of morpheme in the argument	Group N	Group A	Group V	Group I	Group O	Total number of morphemes excluding group O	Total number of morphemes including group O
1	457	61	240	4	15	762	777
2	174	0	6	0	466	180	646
3	93	6	16	0	256	115	371
4	87	4	4	0	139	95	234
5	48	1	3	0	84	52	136
6	28	0	0	0	54	28	82
7	17	0	3	0	38	20	58
8	12	0	0	0	17	12	29
9	3	0	0	0	15	3	18
10	3	0	0	0	6	3	9
11	2	0	0	0	2	2	4
12	0	0	0	0	1	0	1
Total	924	72	272	4	1093	1272	2365

Employing frequency data from Tab. 5 for both cases, including group O and excluding group O, we sorted the amounts of morphemes by descending frequency, and obtained their relative frequencies (p) for each position. We also calculated the negentropy ($-H$) for both cases employing the formula (1) shown above. The negentropy of the number of parts of speech including the group O and excluding the group O are shown in Fig. 8. The relative frequencies for each position in the argument excluding the group O are shown in Tab. 6.

The negentropy of the parts of speech including the group O draws a more complicated curve than the negentropy excluding the group O, and it does not simply decrease while the position of arguments increases from the first to the last of the argument in the clause. Functional words, e.g. postpositions, auxiliary verbs, prefixes and suffixes, which belong to the group O, do not take the first position in the argument and they must follow other words. Therefore, the probability of alternatives for the group O is not different from words from the other groups. We assume the reason may be that words of the group O cause a certain noise in the negentropy in Fig. 8b, and that the negentropy of morphemes should exclude the group O.

The negentropy of the parts of speech excluding the group O, shown as a solid line in Fig. 8, decreases more rapidly than the negentropy of the dependency types in Fig. 5. It can be interpreted that the probability of alternatives of the parts of speech decreases from the second position of the argument or later, if we exclude words of the group O.

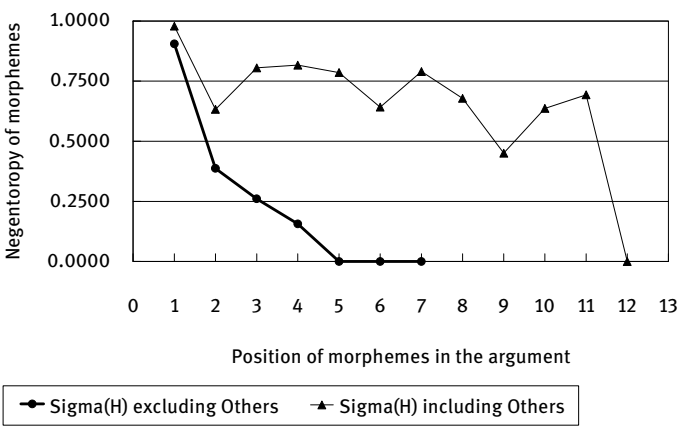


Fig. 8: Negentropy of morphemes and the position of morphemes in the arguments

Tab. 6: Number of parts of speech and their relative frequencies by group (*) and by position in the argument excluding group O and sorted by descending order of frequency (* Groups of the part of speech. N: nouns and pre-nouns. A: adverbs, adjectives and adjective verb. V: verbs. I: interjections and conjunctions. O: others including postpositions, auxiliary verbs, prefixes and suffixes.)

Pos. 1			**Pos. 2**			**Pos. 3**		
Type	*n*	*p*	Type	*n*	*p*	Type	*n*	*p*
N	450	0.5929	N	262	0.8973	N	114	0.9421
V	244	0.3215	V	23	0.0788	A	4	0.0331
A	61	0.0804	A	7	0.0240	V	3	0.0248
I	4	0.0053						
Sigma(*n*)	759		Sigma(*n*)	292		Sigma(*n*)	121	
−Sigma(*H*)		0.9050	−Sigma(*H*)		0.3869	−Sigma(*H*)		0.2605

Pos. 4			**Pos. 5**			**Pos. 6**		
Type	*n*	*p*	Type	*n*	*p*	Type	*n*	*p*
N	53	0.9636	N	26	1.0000	N	12	1.0000
V	2	0.0364						
Sigma(*n*)	55		Sigma(*n*)	26		Sigma(*n*)	12	
−Sigma(*H*)		0.1562	−Sigma(*H*)		0.0000	−Sigma(*H*)		0.0000

Pos. 7		
Type	*n*	*p*
N	7	1.0000
Sigma(*n*)	7	
-Sigma(*H*)		0.0000

5 Discussions on the Argument Order and the Dependency Type

To calculate the negentropy of the dependency types or that of the groups of parts of speech, we employed relative frequencies (*p*) of all data (Tab. 3 and Tab.6). As already mentioned above, we used relative frequencies rather than the number of dependency types or that of groups of parts of speech (i.e. num-

ber of alternatives), because our data shows that relative frequencies are not homogeneous. However, we should also consider the total length of the clause or the total length of the argument if we would like to calculate more exact probabilities of alternatives. The question arises, whether the probability of alternatives for the dependency types in the second argument of the clause, which has three arguments, can be treated equally as the probability of alternatives in the second of eight arguments. For example, in Tab. 3 the relative frequency for the Object is shown as 0.2510, which is calculated from frequencies of Objects in all clauses, and it does not consider the length of clauses. We will explore a solution to obtain the negentropy considering data for the total length in the future.

As a preliminary study of the probability of alternatives by length, we analyzed distributions of dependency types by clause length. Distributions in Fig. 4 employed frequency data of arguments from clauses of various lengths. The number of predicates by position in Tab. 4, i.e. 0, 82, 80 ... and 1, are also distributions of the number of clauses by clause length in arguments because Japanese is a SOV language and the predicate is at the end of the clause. In general, exceptions are possible in rhetorical or colloquial expressions in Japanese. A predicate may not appear at the end of a clause for instance, however, there are no such exceptions in our data. Then we investigated the clauses 80, 54, 19, and 7 with 3, 4, 5, and 6 arguments in the clause by omitting clauses with 1, 2 or 8 arguments (Fig. 9a, 9b, 9c and 9d), because it is difficult to show a significant tendency employing examples with 1 or 2 data points or only one example with 8 arguments in the clause.

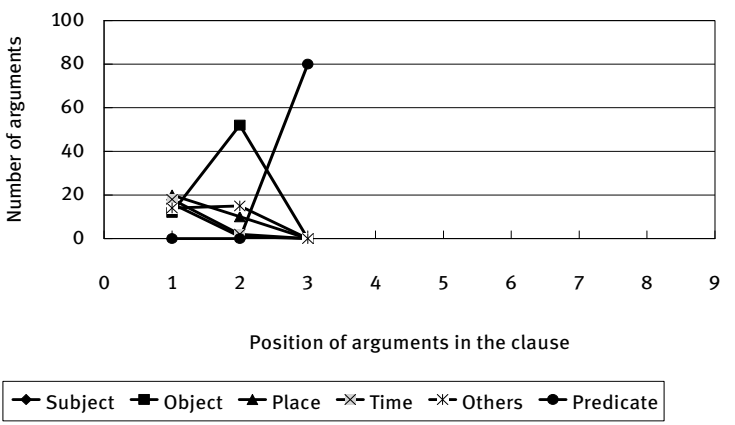


Fig. 9a: Distribution of the number of arguments by type (Clause length =3, *n* = 80 arguments)

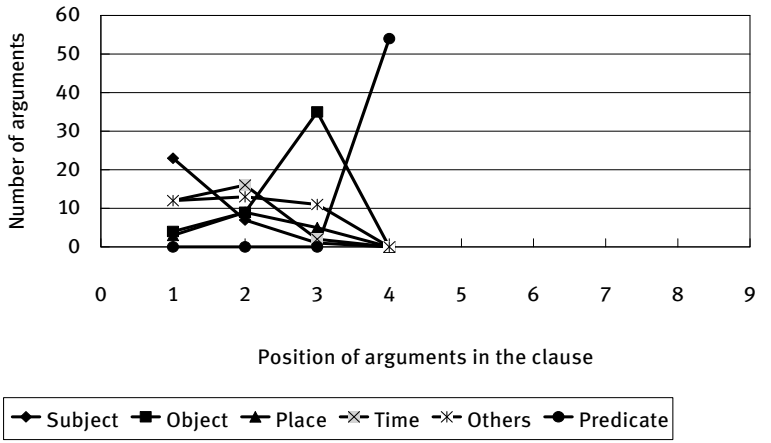


Fig. 9b: Distribution of the number of arguments by type (Clause length =4, *n* = 54 arguments)

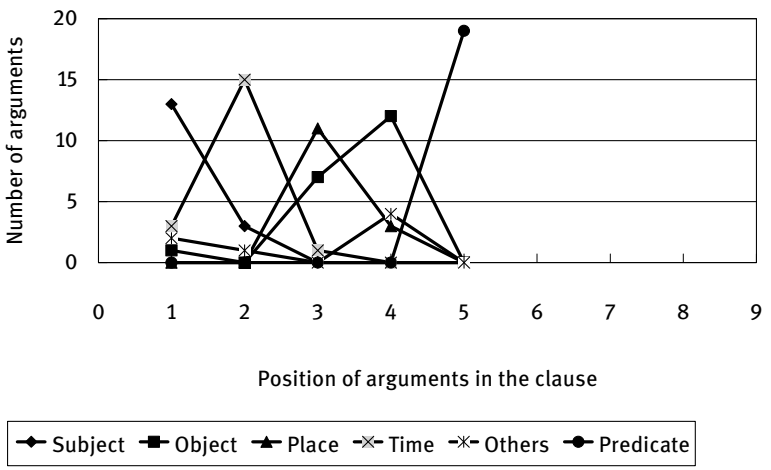


Fig. 9c: Distribution of the number of arguments by type (Clause length =5, *n* = 19 arguments)

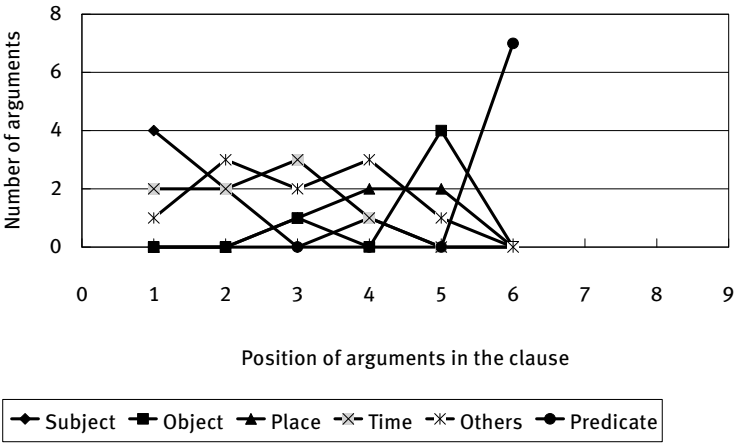


Fig. 9d: Distribution of the number of arguments by type (Clause length = 6, *n*= 7 arguments)

For the verb "meet", it is well known that the standard order of arguments in Japanese is "Subject - Time - Place - Object - Predicate". No canonical word order in Japanese is shown in any grammatical books, except for predicates. However, there are some common word orders which are often shown in textbooks as sample sentences.

By means of observing Fig. 4 and Fig. 9a to 9d, we assume that the predicate may follow the object, that the place may follow the time, and that the predicate may also follow the place. Therefore, the following hypotheses are considered in this section:

- Hypothesis 1. The object appears in a position prior to the predicate.
- Hypothesis 2. The time appears in the position prior to the place.
- Hypothesis 3. The place appears in two positions prior to the predicate.

We investigated all the combinations of dependency types including (1) the object and the predicate, (2) the time and the place, and (3) the place and the predicate for our hypotheses. All data points were taken to obtain functions between positions of two dependency types. We calculated the average of a position if one clause had two or more arguments of the dependency type, such as appositive words "Chu-So kokkyo no machi, Sonkoku" (a city on the border of China and the Soviet Union, Xunke) (JCO0221397). Functions with an average of *y*, such as in Fig. 1, generally show their tendency clearly. However, functions with only two or three averages were not reliable if we omitted averages with less than 10 data points from the regression curve. We obtained a regression line

$y = a * x + b$ for each combination of positions of two dependency types, although regression lines are a special case of regression curves. We can interpret the regression line as follows: the argument y directly follows the argument x if $b = 1$. The argument y always appears with the argument x if $a = 1$. Regression lines and the coefficients of the determination R^2 are shown in Tab. 7. We did not employ regression lines for reverse combinations as we used all data points and the coefficient of the determination R^2 is the same as the original functions. Relationships of six out of ten combinations are also shown in Fig. 10a to 10f. Circles in Fig. 10a to 10f show the number of data points, and black lines show regression lines.

Tab. 7: Regression lines of combinations of positions of the dependency types

x , y	Subject	Object	Place	Time	Others
Object	$y = 0.2001x + 0.5652$. $R^2 = 0.2163$.				
Place	$y = 0.3553x + 0.2763$. $R^2 = 0.2444$	$y = 0.6765x + 1.3425$. $R^2 = 0.4385$. Fig. 10e			
Time	$y = -0.3068x + 2.0114$. $R^2 = 0.0699$.	$y = 0.5999x + 1.9467$. $R^2 = 0.1202$.	$y = 1.1182x + 1.1998$. $R^2 = 0.5363$. Fig. 10f		
Others	$y = -0.2869x + 2.0901$. $R^2 = 0.1324$.	$y = 0.1766x + 2.1024$. $R^2 = 0.0244$.	$y = 0.413x + 1.8696$. $R^2 = 0.0553$.	$y = 0.0397x + 1.706$. $R^2 = 0.0027$.	
Predicate	$y = 0.1945x + 0.464$. $R^2 = 0.1382$. Fig. 10a	$y = 0.8733x - 0.7924$. $R^2 = 0.8095$. Fig. 10b	$y = 0.8206x - 0.983$. $R^2 = 0.787$. Fig. 10c	$y = 0.3916x + 0.0104$. $R^2 = 0.3707$. Fig. 10d	$y = 0.5341x - 0.092$. $R^2 = 0.4214$.

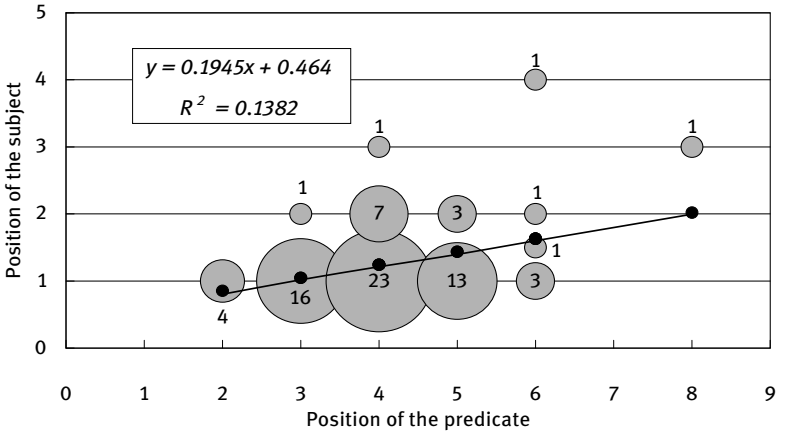


Fig. 10a: The position of the predicate (*x*) and the position of the subject (*y*) in the clause

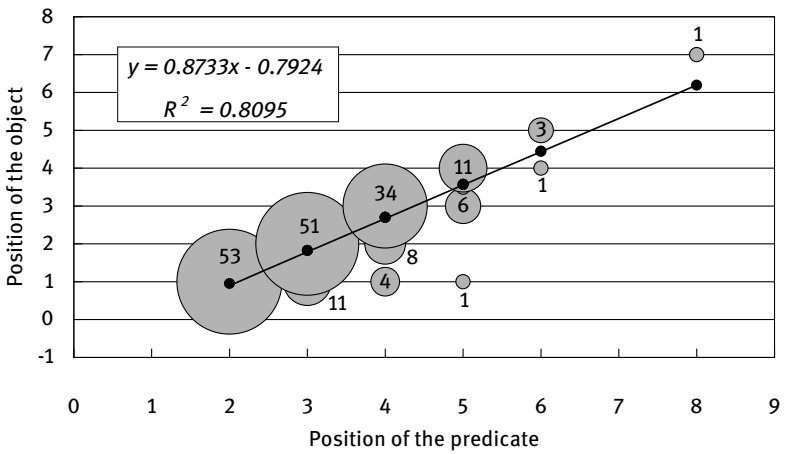


Fig. 10b: The position of the predicate (*x*) and the position of the object (*y*) in the clause

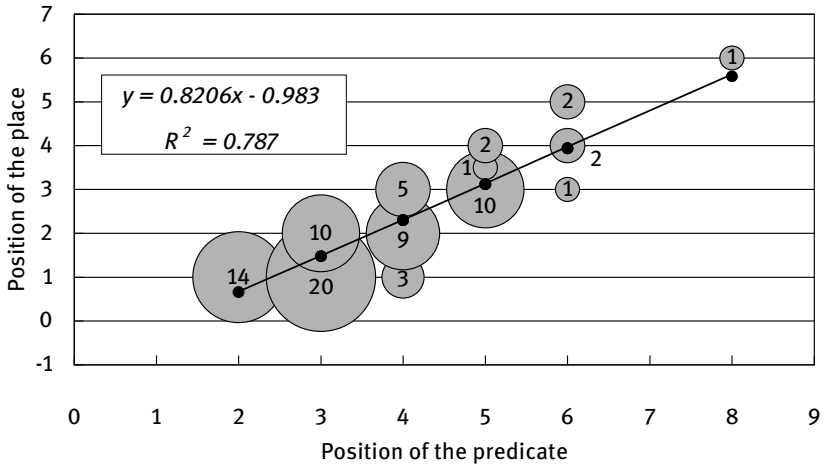


Fig. 10c: The position of the predicate (*x*) and the position of the place (*y*) in the clause

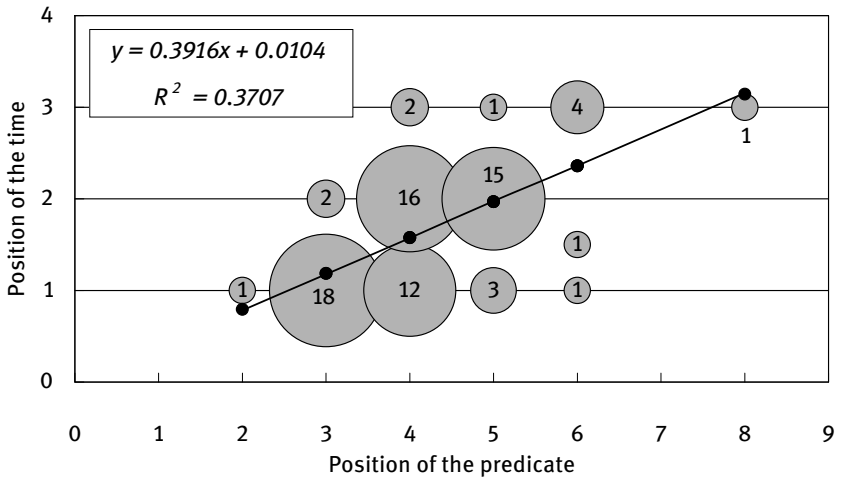


Fig. 10d: The position of the predicate (*x*) and the position of the time (*y*) in the clause

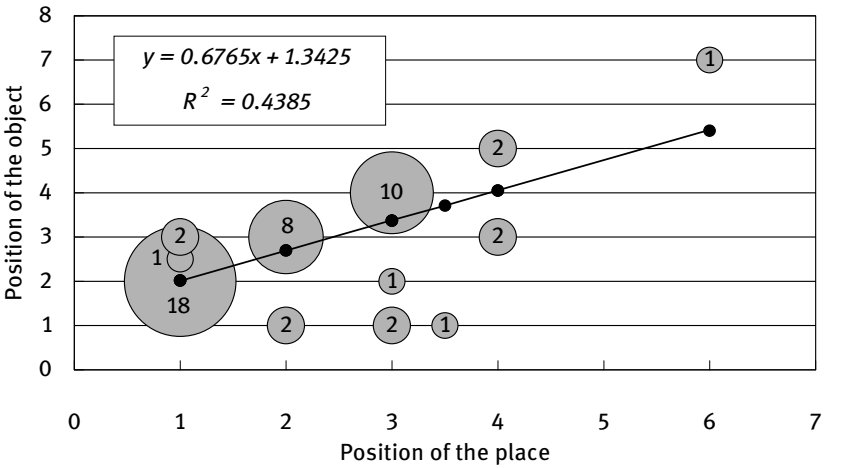


Fig. 10e: The position of the place (*x*) and the position of the object (*y*) in the clause

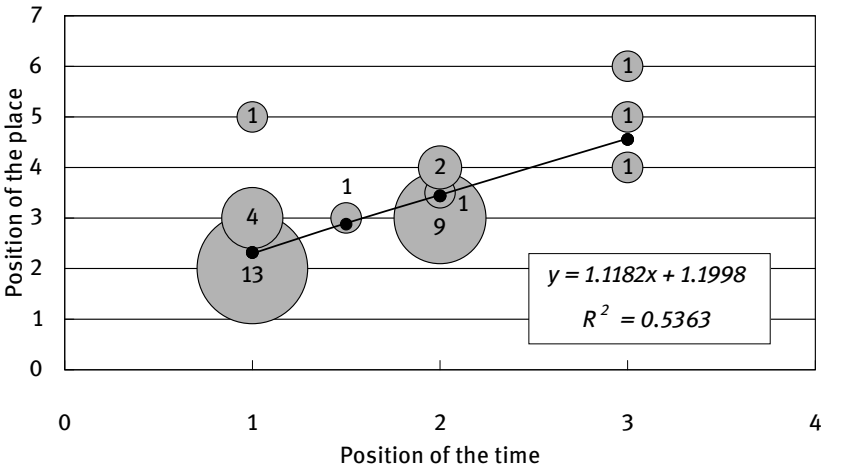


Fig. 10f: The position of the time (*x*) and the position of the place (*y*) in the clause

In Fig. 10a we can observe that most of the data points of the subject take the first position in the clause. This topic should be analyzed in detail in the future.

The relationship between the object and the predicate fits the regression line well with $R^2 = 0.8095$ as shown in Fig. 10b, which is followed by the rela-

tionship between the place and the predicate with R^2 = 0.787 as shown in Fig.10c. The coefficient of determination R^2 for the relationship between the time and the predicate is 0.3707 as shown in Fig. 10d, which is lower than the two relationships mentioned above.

The relationship between the place and the object is shown in Fig. 10e with R^2 = 0.4385. It can be interpreted that the relationships between the object and the predicate, and between the place and the predicate are both relatively strong combinations from the point of view of the position in the clause, and that the relationship between the object and the place is not strong. However, the coefficient *b*, the relationship between the object and the place, is 1.3425, and it shows that there is a tendency for the object to follow the place.

The relationship between the time and the place is shown in Fig. 10f with R^2 = 0.5363. It can be interpreted that the time has a stronger relationship with the place than with the predicate. Fig. 10f also shows all data points are located over the line of $y=x$, and it can be interpreted that the time always appears earlier than the place.

It can be said that Hypotheses (1) and (2) can be accepted, while Hypothesis (3) cannot be directly accepted. From our observations it can be assumed that a clause has three "lumps" of arguments: 1) the subject, 2) the time and the place, and 3) the object and the predicate. We can identify these three "lumps" if the clause is relatively long, and it is difficult to see such "lumps" if the clause is short and arguments are "compressed" in the shorter clause.

6 Conclusions

In the present study we investigated the decrease of the negentropy of dependency types and of parts of speech in the clause. The alternatives for arguments or parts of speech which express information must decrease if the word moves from left to right in the clause. The probability of arguments or parts of speech occurring is not homogeneous in every position of the clause. Therefore, we did not use the amount of different types of words, but rather the relative frequency of words to obtain the negentropy.

The negentropy of the parts of speech excluding the group O decreases more rapidly than the negentropy of the dependency types. Functional words which were included in the group O caused noise in the negentropy value, because the functional words do not have the equal possibility of occurring with a word in the clause as the non-functional words.

Regarding the order of dependency types, we confirmed that a clause has three "lumps" of arguments; i.e. 1) the subject, 2) the time and the place, and 3) the object and the predicate, and that the relationships between the time and the place, and between the object and the predicate are strong from the point of view of the position in the clause. The subject frequently takes the first position, and it does not depend on the length of the clause or the dependency type which follows the subject. The object is frequently followed by the predicate and the time is often followed by the place. We conclude that the appearance of some dependency types do not depend on the absolute position in the clause, but on the relative position to other dependency types, and that these frequent combinations of dependency types or parts of speech can be interpreted by means of effects of the preference of linguistic elements and of the background of the negentropy, i.e. Köhler's register hypothesis (Köhler 2012: 85–86) mentioned in Section 1. It can be assumed that such preference of linguistic elements has an effect on the probability of certain words occurring.

It can be assumed that dependency types and parts of speech have a quantitative relationship because for the subject or the object at least a noun must be taken and for the predicate often a verb is taken.[1] Therefore, choosing a dependency type or choosing a part of speech has a certain effect on each other. This topic will be considered as one of our future tasks.

Acknowledgement: This work is partly supported by the Alexander von Humboldt Foundation [grant number 2014-2016] and the "Grant-in-Aid for Scientific Research (C)" (Project number 16K02741) of the Japan Society for the Promotion of Science (JSPS).

References

Čech, Radek & Ludmila Uhlířová. 2014. Adverbials in Czech: Models for their frequency distribution. In Ludmila Uhlířová, Gabriel Altmann, Radek Čech & Jan Mačutek (Eds.), *Empirical Approaches to Text and Language Analysis* (pp. 45–59). Lüdenscheid: RAM-Verlag.

Helbig, Gerhard & Wolfgang Schenkel. 1969, 1983. *Wörterbuch zur Valenz und Distribution deutscher Verben*. Leipzig: Bibliographisches Institut (Reprint by De Gruyter).

Ishiwata, Toshio & Takano Ogino. 1983. Nihongo Yogen no Ketsugoka (Valency of declinable words in Japanese). In Sh Mizutani (Ed.), *Bunpo to Imi 1* (Grammar and meaning 1) (pp. 226–272). Tokyo: Asakura Shoten.

1 An adjective or a noun with a copula can be also a predicate in Japanese.

Kabashima, Tadao. 1954. Gendaibun ni okeru hinshi no hiritsu to sono zogen no yoin ni tsuite (On proportions of the part of speeches in texts written in the present Japanese and factors which decide the proportion). *Kokugogaku* (Japanese linguistics), *18*, 15–20.

Kabashima, Tadao. 1955. Ruibetsu shita hinshi ni mirauru kisokusei (A regularity in the classified part of speeches). *Kokugo kokubun* (Japanese linguistics and literature), *24* (6), 55–57.

Kabashima, Tadao. 1957. Hyogenron no koso (2) (A concept of the art of writing (2)). *Saikyo daigaku gakujutsu hokoku jinbun* (The Scientific reports of Saikyo University, Humanistic science), *10*, 23–48.

Köhler, Reinhard. 1999. Syntactic structures: Properties and interrelations. *Journal of Quantitative Linguistics*, *6* (1), 46–57.

Köhler, Reinhard. 2012. *Quantitative Syntax Analysis.* Berlin: Mouton De Gruyter.

Minami, Fujio. 1974. *Gendai nihongo no kozo* (The structure of the present Japanese). Tokyo: Taishukan.

Minami, Fujio. 1993. *Gendai nihongo bunpo no rinkaku* (The outline of the grammar of the present Japanese). Tokyo: Taishukan.

National Language Research Institute. 1964. *Gendai Zasshi 90shu no Yogo Yoji: Dai3bunsatsu: Bunseki* (Vocabulary and Chinese Characters in Ninety Magazines of Today: vol.3: Analysis of Results). Tokyo: Shuei Shuppan.

Ogino, Takano, Masahiro Kobayashi & Hitoshi Isahara. 2003. *Nihongo Doshi no Ketsugoka* (Verb valency in Japanese). Tokyo: Sanseido.

Sanada, Haruko. 2012. Joshi no Shiyo Dosu to Ketsugoka ni Kansuru Keiyoteki Bunseki Hoho no Kento (Quantitative approach to frequency data of Japanese postpositions and valency). *Rissho Daigaku Keizaigaku Kiho* (The quarterly report of economics of Rissho University), *62* (2), 1–35.

Sanada, Haruko. 2014. The choice of postpositions of the subject and the ellipsis of the subject in Japanese. In Ludmila Uhlířová, Gabriel Altmann, Radek Čech & Jan Mačutek (Eds.), *Empirical Approaches to Text and Language Analysis* (pp. 190–206). Lüdenschied: RAM-Verlag.

Sanada, Haruko. 2015. A co-occurrence and an order of valency in Japanese sentences. In Arjuna Tuzzi, Jan Mačutek & Martina Benešová (Eds.), *Recent Contributions to Quantitative Linguistics* (pp. 139–152). Berlin: Walter de Gruyter.

Sanada, Haruko. 2016a. Menzerath-Altmann law and the sentence structure. *Journal of Quantitative Linguistics*, *23* (3), 256–277.

Sanada, Haruko. 2016b. Kabashima Funktion. In Reinhard Köhler, Peter Grzybek & Sven Naumann (Eds.), *Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK)*, vol.9. *Quantitative und Formale Linguistik* (German Edition). Online Version. Berlin, New York: Walter de Gruyter.

Sanada, Haruko. 2016c. A measurement of the part of speech in the text using the noun-based proportion. In Emmerich Kelih, Roisin Kight, Jan Macutek & Andrew Wilson (Eds.), *Issues in Quantitative Linguistics* 4 (*Studies in Quantitative Linguistics*, vol. 23) (pp. 82–93). Lüdenscheid: RAM-Verlag.

Sanada, Haruko. 2018a. Quantitative aspects of the clause: Length, position and depth of the clause. *Journal of Quantitative Linguistics.* DOI: 10.1080/09296174.2018.1491749

Sanada, Haruko. 2018b. Quantitative interrelations of properties of complement and adjunct. In Lu Wang, Reinhard Köhler & Arjuna Tuzzi (Eds.), *Structure, Function and Process in Text* (pp.78–99). Lüdenscheid: RAM-Verlag.

Sanada, Haruko. To appear. Length of clauses and a perspective on the three dimensional model of Synergetic Linguistics. *Journal of Quantitative Linguistics.*
Tesnière, Lucien. 1959, 1988. *Éléments de Syntaxe Structurale. 2nd edition.* Paris: Klincksieck.

Software

Graduate Schools of Informatics in Kyoto University; NTT Communication Science Laboratories. 2008. Morphological analyzer: *MeCab*, version 0.97. (https://code.google.com/p/mecab/)
National Institute for Japanese Language and Linguistics. 2008. Digital dictionary for the natural language processing: *UniDic*, version 1.3.9. (http://www.ninjal.ac.jp/corpus_center/unidic/)

Qian Lu, Yanni Lin, Haitao Liu*

Dynamic Valency and Dependency Distance

Abstract: Dependency is dynamic manifestation of valency, and dependency distance (DD) is closely related to syntactic structure. By introducing the concept *degree* in graph theory, we analyze the relationship between dynamic valency (DV) and DD based on the Chinese and English treebanks. Our findings are: (1) the mean dependency distance (MDD) of the Chinese treebank is greater than that of English, while the variance of DV of Chinese is lower than that of English; (2) some values of the variance of DV exist in Chinese but not in English; (3) at a specific sentence length, there is a linear relationship between MDD and the variance of DV in the syntactic structures in English and Chinese. These findings suggest: (1) Chinese may have some unique syntactic dependency structures that are not found in English; (2) high DV may contribute to high MDD, but this effect on MDD may not be stronger than some grammatical factors.

Keywords: dependency distance; dependency grammar; dependency treebank; syntactic complexity; valency; dynamic valency

1 Introduction

Recently, analyzing the linguistic features by using authentic language materials has become one of the most important approaches to exploring the laws in human languages. Both mathematical deduction and treebank-based statistical studies have revealed that the dependency distance between two syntactically related words in a linear order tends to be minimized (Liu 2007 2008a; Temperley 2008; Ferrer-i-Cancho 2004 2006 2013 2014; Futrell et al. 2015; Liu et al. 2017). Dependency distance minimization, to a certain extent, shapes the syntactic patterns of human languages (Liu et al. 2017). So, what role do some features of syntactic structures play in dependency distance minimization? Do they

Qian Lu, Guangxi Normal University, Guilin, P.R.China
Yanni, Lin, Zhejiang University, Hangzhou, P.R.China
Haitao Liu, Zhejiang University, Hangzhou; Guangdong University of Foreign Studies, Guangzhou, P.R.China, htliu@163.com

https://doi.org/10.1515/9783110573565-008

reflect the commonalities and differences in human languages? These questions have aroused the interest from disciplines of quantitative linguistics, psycholinguistics and cognitive science (Liu et al. 2017).

The universal tendency of dependency distance minimization in human languages is assumed to result from the commonality of human cognitive mechanisms (Ferrer-i-Cancho 2004; Liu 2008a; Jiang & Liu 2015). Studies based on large-scale corpora show that there is a threshold that the MDD of a natural language does not exceed (Liu 2008a), which is constrained by human cognitive capacity (Ferrer-i-Cancho 2004 2006 2013 2014; Liu 2008a; Jiang & Liu 2015). Specifically, human working memory is similar and has an estimated capacity of about 4 (Miller 1956; Cowan 2001), whereas human syntactic processing mechanism is based on the incremental strategy (Liu 2008a; Jiang & Liu 2015). Working memory capacity or cognitive mechanism restrains the mean dependence distance (MDD) between words in a linear order less than 4 in a sentence (Liu 2008a; Jiang & Liu 2015). Thus, if the MDD is greater than 4, more computational cost on sentence processing will rise. That is, this will lead to difficulty in language understanding, violating the least effort principle (Zipf 1949). However, it is interesting that human languages, while characterized by universality such as dependence distance minimization, manifest their structural diversity at the same time. In this sense, analyzing the structural features of dependency trees and their motivation of formation so as to detect the patterns across languages may shed light on how human cognitive mechanism affects human languages.

Using authentic language treebanks, scholars have studied structural features related to dependence distance. Early researchers considered projectivity has an impact on dependency (Lecerf 1960; Hays 1964), namely the non-projective structure can significantly increase the MDD of a language. Further studies have found that projective structure is more likely to be a by-product of dependency distance minimization (Ferrer-i-Cancho 2006 2013 2014), rather than vice versa. In addition, chunking plays an important role in reducing dependency distance. Chunking, together with projectivity, can reduce the MDD of a random language to the scope of natural language (Lu et al. 2016). Besides, despite apparent language varieties, human languages develop the regularity of word orders in a way that helps restrain the possible increase in MDD caused by long-distance dependencies (Gildea & Temperley 2010; Liu et al. 2017). The previous studies have displayed a complicated picture of the relation between dependency distance and syntactic dependency structure, with various structural features being in a state of competition and co-existence in a language system.

The structural features of dependency trees mentioned above are considered to have a certain relationship with the valency of words in a sentence. Valency is originally defined as "the ability of a verb to open up certain positions in its syntactic environment, which can be filled by obligatory or optional complements" (Herbst 2007). In a broader sense, it refers to the potential ability of a word to combine with other words (Liu 2011), which is materialized from the covert, static state in the dictionary to the overt, dynamic state to form a sentence structure (Liu 2009a 2009b). Dynamic valency of a word can be roughly measured by degree, namely the number of its connections (Liu 2008b; Zhang & Liu 2017). Therefore, dynamic valency is related to MDD for three main reasons. Firstly, the 2nd moment of degree (dynamic valency) of a tree (written as $\langle k^2 \rangle$, where k is the degree of the node (word)) determines the minimum value of MDD (Ferrer-i-Cancho 2013). In other words, $\langle k^2 \rangle$ must be reduced so that the nodes with higher dynamic valencies in the tree can be avoided, resulting in a smaller MDD. Secondly, the number of occurrences of dependency crossing structures is bounded by $\langle k^2 \rangle$ (Ferrer-i-Cancho 2013). Further research has found that when $\langle k^2 \rangle$ is maximized, crossings will no longer exist, and the expected value of $\langle k^2 \rangle$ is only related to sentence length (Ferrer-i-Cancho 2014). But whether this finding suits authentic language material remains to be investigated. Finally, if dynamic valencies of words are similarly distributed in several syntactic dependency trees (namely the identical values of $\langle k^2 \rangle$), the size and position of chunks will affect MDD (Lu et al. 2016).

Most of the theoretical assumptions above focus on the relationship between degree of words and dependency distance in random dependency trees. Based on the authentic treebanks of two languages, we attempt to explore the role of dynamic valency, i.e., whether valency, as a potential motivation for forming dependency, has a certain relationship with dependency distance in a natural language. In order to answer this major question, we will embark on the following more concrete questions based on Chinese and English dependency treebanks:

(1) Is the mean dependency distance of Chinese sentences with various lengths different from that of English? Does it change with sentence length?
(2) Is the variance of dynamic valency of Chinese sentences with various lengths different from that of English? If yes, what are the reasons?
(3) Do the relationships between the variance of dynamic valency and MDD differ in two languages? If yes, what are the reasons for the differences?

The first question is to verify whether the statistical results of dependency distance based on treebanks accord with the conclusions of previous studies, while the rest are designed to explore the possible role of dynamic valency in dependency distance minimization as well as its similarities and differences in two languages.

The rest of this paper is organized as follows. Section 2 introduces the research methods and materials. Section 3 presents the major results and discussions on MDD and the variance of dynamic valency in Chinese and English as well as the relationships between them; the special sentence structures influencing dynamic valency and dependency distance in both languages are also analysed. Concluding remarks and suggestions for further research are given in the last section.

2 Methods and Materials

Dependency distance (DD) is measured by the number of intervening words between two syntactically related words of a sentence in a linear order (Hudson 1995; Liu 2007 2008a). Originated from syntactic research (Heringer et al. 1980; Liu et al. 2017), it is now used in a series of studies within the frameworks of dependency grammar and phrase structure grammar. DD reflects the cognitive mechanism of human beings in language production and understanding (Liu 2007 2008a; Ferrer-i-Cancho 2004 2006 2013 2014; Jiang & Liu 2015; Lu et al. 2016). In this study, based on the theory of dependency syntax (Hudson 2010; Liu 2009a; Mel'čuk 1988), we attempt to probe the relationship between dependency distance and dynamic valency of words in a sentence. In terms of the measurement of dependency distance, what Liu (2007 2008a) calculates is the mean dependency distance (MDD), while Temperley (2007) and Futrell et al. (2015) adopt another term "dependency length" and take the mean of total dependency length of sentences (MDL) as the metric. The main difference between the two methods lies in whether to take sentence length into account. The former method, which eliminates the impact of sentence length on dependence distance, is favoured by us in comparative analysis of different languages. In our study, the dependency distance with the sentence length of 1 is normalized to 1 for the sake of convenience in calculation.

In Fig. 1, the sentence root "*like*" is governed by a fake node "ROOT"; the arc connecting a pair of syntactically related words represents a dependency relation; the label above this arc denotes the type of dependency (i.e., subject (SUBJ), object (OBJ), and attribute (ATR)), and the number below is the depend-

ency distance. So the MDD of the sentence "*I like red apples*" is: (1+2+1)/3 = 1.33. This method of calculation applies to a sentence, a text or a corpus (Liu 2007 2008a).

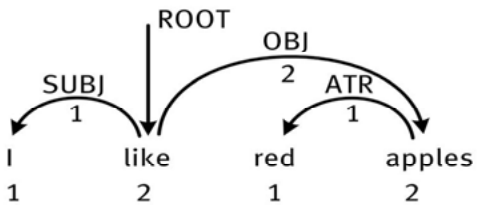


Fig. 1: The dependency tree of "*I like red apples*"

When a sentence is produced, the valency of a word turns from recessive state to dominant state materialized through dependency relation (Liu 2009a 2009b). In other words, the static valency in the dictionary becomes dynamic valency during the linearization of sequence. The dynamic valency of a word can be roughly measured by the degree of node in the dependency tree, including in-degree (governed) and out-degree (governing) (Liu 2008b), which refers to the number of edges incident with it in a graph (or tree) (Rosen 2012).

The last line in Fig. 1 provides the dynamic valency of each node. The dynamic valency of root "*like*" is 2, the root (in-degree) not being taken into consideration. The variance of the dynamic valency (written as $V[k]$) can be calculated by $\langle k \rangle$ and $\langle k^2 \rangle$ (k is the dynamic valency of a word).

$\langle k \rangle$ and $\langle k^2 \rangle$ of the dynamic valency (degree) mentioned above can be calculated according to the following equation (Rosen 2012; Ferrer-i-Cancho 2013):

$$\langle k \rangle = 2 - 2/n \tag{1}$$

$$\langle k^2 \rangle = \sum k_i^2/n,\ i = 1, \ldots, n \tag{2}$$

where n is sentence length. Accordingly, as for the sentence "*I like red apples*", $\langle k \rangle = 2 - 2/4 = 1.5$, $\langle k^2 \rangle = (1^2+2^2+1^2+2^2)/4 = 2.5$.

$V[k]$ in a single sentence is calculated by the difference between $\langle k \rangle^2$ and $\langle k^2 \rangle$. As for a set of sentences or a text, the variance of dynamic valency in a language is the mean of total $V[k]$ of all the sentences.

$$V[k] = \langle k^2\rangle - \langle k\rangle^2 \tag{3}$$

$$\langle V[k]\rangle = \sum V[k]/sn = \sum(\langle k^2\rangle - \langle k\rangle^2)/sn \tag{4}$$

where sn is the number of sentences. For the sentence in Fig. 1, $V[k] = 2.5-1.5^2 = 0.25$ according to equation (3).

It can be deduced from equation (1) - (4) that $\langle k\rangle^2$ is related to sentence length n only, and hence $V[k]$ and $\langle V[k]\rangle$ are related to $\langle k^2\rangle$ only. So, $V[k]$ of different sentences is actually captured by $\langle k^2\rangle$. In our study aimed to compare the relationships between degree and dependency distance at different sentence lengths, we need to figure out $\langle k\rangle^2$ so as to get the $\langle V[k]\rangle$ of each language.

In addition, $\langle V[k]\rangle$ in random dependency trees can be mathematically derived in reference to the related work (Ferrer-i-Cancho 2013 2014), i.e., the expected value of $\langle k^2\rangle$ in random dependency tree is:

$$E[\langle k^2\rangle] = (1-1/n)(5-6/n) \tag{5}$$

Since $\langle k\rangle$ is related to the sentence length only, the expected value of $\langle V[k]\rangle$ of random dependency tree is:

$$E[\langle V[k]\rangle] = E[V[k]] = (n-1)(n-2)/n^2 = 1-3/n+2/n^2 \tag{6}$$

In this study, the English and Chinese dependency treebanks are used to investigate the interrelationships between sentence length, dependency distance and syntactic structures. As some indicators may be language-specific, it seems more objective to examine the relationship with the treebanks of more than one language.

The Chinese dependency treebank used in our study (Qiu et al. 2014) contains the articles collected from *the People's Daily* from January 1, 1998 to January 10, 1998 (the news genre). There are altogether 14,463 sentences and 336,138 words (punctuation marks included). Our English dependency treebank (Kato et al. 2016) is converted from the Penn Treebank (annotated in phrase structure grammar), which was originally aimed to accomplish the task of recognizing multiword expressions in natural language processing. During the conversion, the constructions violating some basic rules in dependency grammar such as multi-dominators and circuits are avoided. Sourced from Section 00-24 of *Wall Street Journal* (the news genre), it has a total of 37,015 sentences and 901,673 words (including punctuation marks).

As dependency reflects the syntactic connection between a pair of words, the actual value of MDD can be distorted if the punctuation marks are not properly handled, since the punctuation marks will increase the MDDs of two languages, as the statistics suggest in the next section. So a further processing is needed:

a. Delete the punctuation mark if it is a dependent, but retain it and then normalize the dependency distance to 1 if it is not.
b. If the punctuation is deleted as stated in the previous step, the words are renumbered consecutively, and then revise the ordinal numbers of their corresponding heads.

Moreover, to ensure the adequate number of sample sentences of all lengths, we made sure the minimum amount of sentences is 38 for each sentence length (SL) ranging from 1 to 50. Fig. 2 shows the distribution of sentence length in the treebanks after the punctuation marks are processed. When the sentence length is larger than 50, the frequency of sentences of some lengths starts to gradually decrease to less than 38. In order to clearly display the relationship between sentence length and frequency, the y-axis in Fig. 2 is set as logarithmic coordinates.

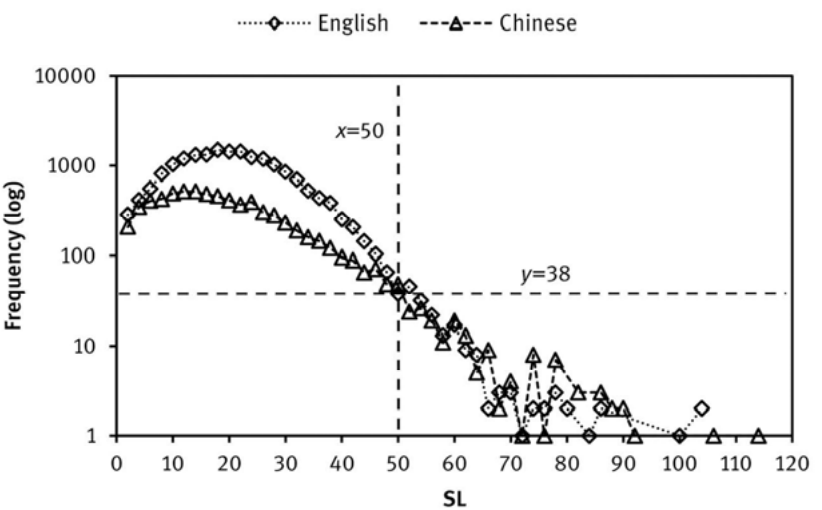


Fig. 2: The distribution of even sentence length in the English and Chinese dependency treebanks (processed)

3 Results and Discussions

3.1 Statistical analysis of MDD based on the English and Chinese Treebanks

Preliminary results show that the dependencies involving punctuation account for 16.28% of the total in the Chinese treebank (the primitive treebank contains 14,463 sentences), and the MDD of the dependencies involving punctuation (5.91) is much higher than that of the treebank (3.79). In the English dependency treebank (the primitive treebank contains 37,015 sentences), the amount of dependencies with punctuation occupies 13.09% of the total, and the MDD of the dependencies involving punctuation (8.45) is much higher than that of the treebank (3.62) as well.

The processed Chinese dependency treebank without punctuation (including 14,457 sentences) has an MDD of 3.06, and the mean sentence length is 19.63. Meanwhile, the English dependency treebank after processing has no change in the total number of sentences, with an MDD of 2.69 and a mean sentence length of 21.3. Chinese always has a higher MDD than English with or without punctuation.

The language material used above is composed of sentences with various sentence lengths. Ferrer-i-Cancho & Liu (2014) argue that mixed sentence lengths may render miscalculation of MDD. Nevertheless, Jiang & Liu (2015) consider that the mixture does not substantially influence the comparative analysis of MDDs in different languages. Our current study is in line with the latter point of view. The results of our statistical analysis of MDDs with different sentence lengths (ranging from 1 to 50) in the English and Chinese treebanks are given in Fig. 3.

As shown in Fig. 3, the MDDs of English and Chinese go up smoothly after all the sentences longer than 50 are removed. When the sentence length falls within the interval of [1, 8], the MDDs in both languages are almost the same. However, when the sentence length reaches more than 8, the MDD of Chinese becomes higher than that of English. The main reason may be that the ratio of adjacent dependencies of different sentence length, though fluctuating within a certain range for English and Chinese (Jiang & Liu 2015), is not enough to cause MDDs to fluctuate violently with the change of sentence length. Moreover, longer sentences lead to an increase in the likelihood of longer dependency relation. But due to the constraints of human cognitive mechanism, the MDDs increase

very slowly with the sentence lengths. In a way, for two languages, the difference in the MDD of the treebanks (mixed sentence lengths) is also reflected in the sentence sets of varying length.

Then a question naturally comes to our mind: Why is the MDD of Chinese always higher than that of English? Recent studies show that the MDD of human language is related to human cognition mechanisms. Sentences with smaller MDDs are easier to process as they facilitate the economy of computational cost (Liu 2008a; Temperley 2008; Ferrer-i-Cancho 2004; Jiang & Liu 2015), observing the least effort principle (Zipf 1949). In this sense, can we deduce that Chinese sentences are more difficult to process than English at the syntactic level? Liu (2008a) proposes the question, and explains the reasons for the higher MDD in Chinese sentences from several aspects including the proportion of adjacent dependencies, syntactic structure, and so on. In another study, Jiang & Liu (2015) use parallel corpora to study the relationship between sentence length and structural features including the proportion of adjacent dependencies, MDD and dependency direction. The greater MDD may indicate a better capacity of Chinese native speakers in processing long dependencies.

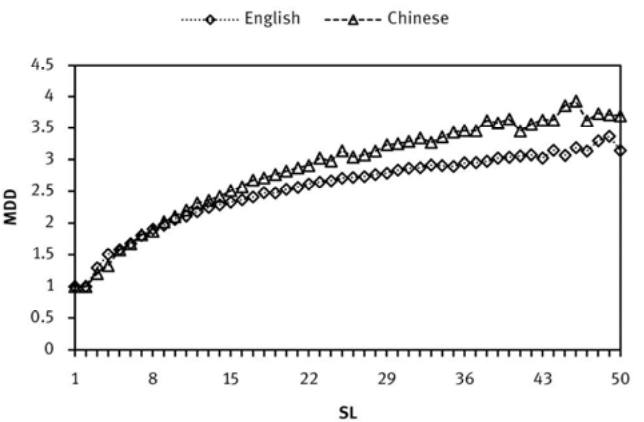


Fig. 3: MDDs in the English and Chinese treebanks by sentence length

In the existing studies, there is a high proportion of adjacent dependencies in 20 languages, more than 50% of which have dependency distance of 1 and 2 (Liu 2008a; Jiang & Liu 2015). This is the result of the interaction between dependency distance minimization and other factors (grammar, hierarchy in dependency

tree, etc.). More attention is supposed to be paid to these factors which influence the syntactic structure patterns of human languages.

Special syntactic structure is another factor influencing dependency distance (Liu 2008a). A prepositional phrase generally follows the noun it modifies in English, while precedes the noun in Chinese, so the complements of a preposition in Chinese will lengthen the dependency distance between the noun and its dependent. In addition, some of the syntactic functions in Chinese differ from English in the way of realization. In perfective aspect, for example, functional words are attached to verbs as heads in Chinese. This also leads to an increase in Chinese MDD, as shown in Fig. 4.

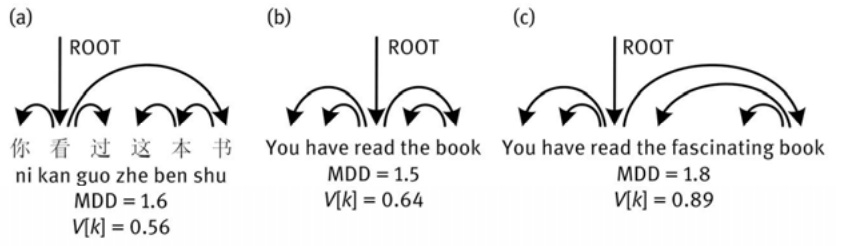


Fig. 4: The dependency trees of three sentences: "ni kan guo zhe ben shu", "You have read the book", and "You have read the fascinating book"

For the Chinese sentence (a) "*ni* (*you*) *kan* (*read*) *guo* (*perfective aspect marker*) *zhe* (*this*) *ben* (*quantifier*) *shu* (*book*)", where "*guo*" marks the perfective aspect, "*zhe*" is a pronoun and "*ben*" is a quantifier. Thus the MDD of the sentence increases to 1.6, which is caused by "*guo*" and "*ben*". In sentence (b), its semantic equivalent in English, "*You have read the book*", the perfective aspect is realized by the auxiliary verb and the past participle, and there is no corresponding quantifier. All this reduces the MDD to 1.5, smaller than Chinese. Admittedly, it remains to be testified whether the assumption in this case applies to more data.

Interestingly, in sentence (c), when the number of words (dynamic valency) connecting to "*book*" in Fig. 4 increases from 2 to 3, the variance of the dynamic valency of the entire sentence grows with the MDD, rising from 1.5 to 1.8. Then, can we infer that the dynamic display of the valency causes an increase in MDD? Or does the rise of MDD merely result from sentence length? In addition, sentences (a) and (c) in Fig. 4 show that MDD and $V[k]$ of the Chinese sentence

are smaller, but it seems futile to use the values of $V[k]$ to explain why the MDD of Chinese is greater than that of English with the same sentence length. So what role does dynamic valency play in both languages?

To answer these questions, we will investigate the relationship between dynamic valency and MDD from two aspects: 1) to examine the variance of dynamic valency ($V[k]$) in both languages (Section 3.2); 2) to examine the relationship between dynamic valency and dependency structure (Section 3.3); 3) to examine the relationship between $V[k]$ and MDD at a specific sentence length (Section 3.4) by statistical analysis.

3.2 Statistical analysis of dynamic valency based on the English and Chinese Treebanks

According to the equation in Section 2, we figured out the mean value of the variance of dynamic valency ($\langle V[k] \rangle$) is 1.15 in Chinese and 1.64 in English. As far as the whole language system is concerned, Chinese may have smaller $\langle V[k] \rangle$ than English, i.e., words with high dynamic valency are less likely to appear in sentences.

Degree (dynamic valency) may reflect the importance of a word in the syntactic structure (Liu et al. 2017). The relationship between degree and syntactic structure is mainly discussed in language networks research (Liu 2008b). In order to explore the interrelations among degree, dependency distance and dependency crossing, Ferrer-i-Cancho (2013) proposes that there may be a close relationship between degree and the minimum value of MDD from the perspective of mathematical theory. He proves that the minimum value of MDD is constrained by the variance of the node degree, namely, the greater the variance of the node degree, the greater the minimum value of MDD.

Since dependency distance in human languages tends to be minimized, $\langle k^2 \rangle$ must be reduced, i.e., the occurrences of the nodes with high dynamic valencies in syntactic trees must be avoided if possible. To specify, high $\langle k^2 \rangle$ indicates the possible existence of several nodes with greater dynamic valencies in the dependency tree. These nodes, with a dependency distance greater than 1, lead to a decrease in the proportion of adjacent dependency and thus an increase in MDD.

However, the data calculated based on the whole treebanks contradicts with the tentative assumption mentioned above that “Chinese with a greater MDD seems to have more words with high dynamic valencies compared with Eng-

lish". It also runs counter to the mathematical deduction of the relationship between $V[k]$ and MDD in the previous studies (Ferrer-i-Cancho 2013; Liu et al. 2017).

Similarly, in order to avoid the impact of the mean variance of dynamic valency ($\langle V[k] \rangle$) induced by mixed sentence lengths, we calculated $\langle V[k] \rangle$ of the sentences at different lengths in two natural languages as well as an artificial random language for better comparison, as shown in Fig. 5.

Fig. 5 shows that as the sentence length grows, the mean variance of dynamic valency ($\langle V[k] \rangle$) in each language gradually increases. Here, we can see that the longer the sentence length, the greater the MDD and $\langle V[k] \rangle$. The results are in line with the relevant inference in the related literature (Ferrer-i-Cancho 2013; Liu et al. 2017). In addition, we also discover that as the sentence length grows, the value of $\langle V[k] \rangle$ in the random dependency tree rises close to 1, which can also be obtained by calculating the limit of the right-hand side of the equation (6) (See Section 2). However, it is noteworthy that when the sentence length is greater than 3, $\langle V[k] \rangle$ of the English dependency tree outnumbers that of Chinese, indicating that larger MDD does not bring greater $\langle V[k] \rangle$.

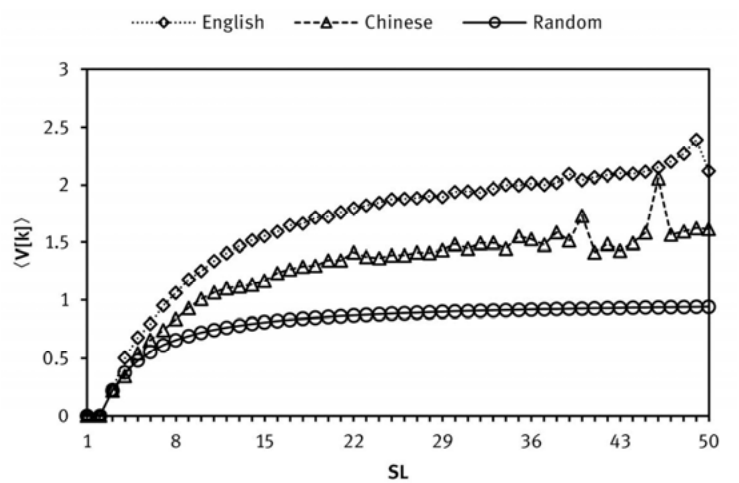


Fig. 5: $\langle V[k] \rangle$ in the English, Chinese and Random language treebanks by sentence length

Random languages are reported to have a larger MDD than natural languages in previous studies (Ferrer-i-Cancho 2004 2006; Liu 2007 2008a; Futrell et al. 2015; Lu et al. 2016). However, our data suggests that $V[k]$ of the random language is the smallest. This could be illustrated by essential difference in the two kinds of

languages. Deemed as a human-driven complex adaptive system (Liu 2014), a natural language is constrained by human cognitive mechanisms and the least effort principle. In this sense, dependency distance minimization shapes the syntactic patterns of natural languages (Liu et al. 2017). By contrast, the probability of the connection between words in random languages is equal, free of human influence. Therefore, it may not make any sense to infer the relationship between $V[k]$ and MDD based on random trees. Given the maximum MDD and minimum $V[k]$ in random trees, it seems somewhat hasty to reach such a conclusion that smaller $V[k]$ in Chinese means larger MDD.

Then, why is $\langle V[k] \rangle$ of Chinese dependency tree no greater than that of English? Can the unique syntactic structures (Liu 2008a) serve as an explanation of smaller dynamic valency but larger MDD in Chinese?

3.3 The relationship between dynamic valency and dependency structure

To answer the question in the previous section, we conducted statistical analysis of $V[k]$ to see whether both languages differ in syntactic structure. Sentences with the length of 19 are selected as material to guarantee the sufficiency in number, and we found 478 Chinese sentences and 1375 English sentences. The results are shown in Fig. 6.

Fig. 6 shows that Chinese and English share many values of $V[k]$ (the variance of the dynamic valency) at a sentence length of 19 except in two special cases: Chinese has the smallest $V[k]$ in the left side of the figure as pointed by the arrow, while English has higher $V[k]$ in the right side.

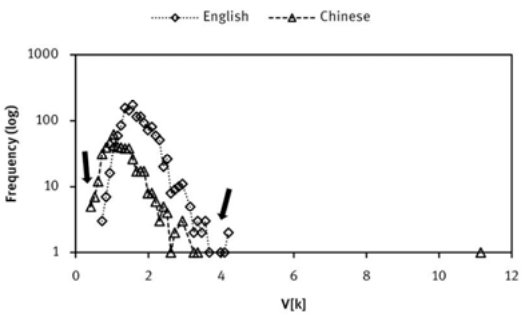


Fig. 6: The relationship between $V[k]$ and its frequency in the English and Chinese treebanks (SL = 19)

The unique values and frequencies of $V[k]$ in both languages at the sentence length of 19 are displayed in Tab. 1. Overall, the unique values of $V[k]$ in English are higher with lower frequencies of occurrence, while those in Chinese are lower but appear more frequently. Fig. 6 shows that, if Chinese has the same number of sentences as English in our dataset, according to the frequency distribution, English may lose the unique value of $V[k]$ in Tab. 1. However, when the number of English sentences is nearly 3 times larger than that of Chinese sentences in our case, Chinese still has some values of $V[k]$ which do not exist in English. Therefore, the dependency trees with smaller $V[k]$ may only appear in Chinese. In other words, Chinese may have particular syntactic structures. For this reason, the dependency trees with smaller $V[k]$ are unique in Chinese. In other words, Chinese has unique syntactic structures.

Tab. 1: The unique values of $V[k]$ and their frequencies in Chinese and English (SL = 19)

Chinese		English	
$V[k]$	Frequency	$V[k]$	Frequency
0.409972	5	2.831025	10
0.515235	7	3.146814	5
0.620499	12	3.462604	2
11.146810	1	3.567867	3
		3.673130	1
		3.988920	1
		4.094183	1
		4.199446	2

In order to verify this tentative assumption, we extend our calculation to the sentences with the lengths from 13 to 21 in both languages. Similar cases are found in both languages.

After a further analysis of the unique sentences in Chinese (i.e. with lower dynamic valencies) as shown in Fig. 7(a) (for more sentences, please refer to the Appendix), the following characteristics are summarized: (1) In the sentence, only the heads of subjects, predicates or objects in the main or small clauses have a degree of no more than 2, the majority of words with a degree of 1 and a minority of 2; (2) Generally, the head of object is placed at the end of the sentence or clause, and the modifier of head as object or subject is longer. If the dependency structure contains multiple nested structures, the MDD will in-

crease rapidly, as shown by Sentence 4 in Appendix (a). The structure in the small clause is similar to that in the main clause.

The pre-noun modifiers in Chinese sentences are longer, while in English sentences there are more evenly distributed words with higher dynamic valencies (with a degree of no less than 3), i.e., a head takes more parallel dependents, as shown in Fig. 7 (b). When the sentence length is 16-32 and the size of the chunk is 4-7, the MDD tends to be smaller (Lu et al. 2016). In spite of more words with higher dynamic valency in English, these words are likely to divide the sentence into multiple regional chunks, thus reducing the MDD. This may partly explain why MDD still remains small in English. According to Fig. 6 and Fig. 7, the special sentence patterns shown above may lead to a smaller $V[k]$ but a larger MDD in Chinese. This is also the reason why English and Chinese show such a great difference in $V[k]$.

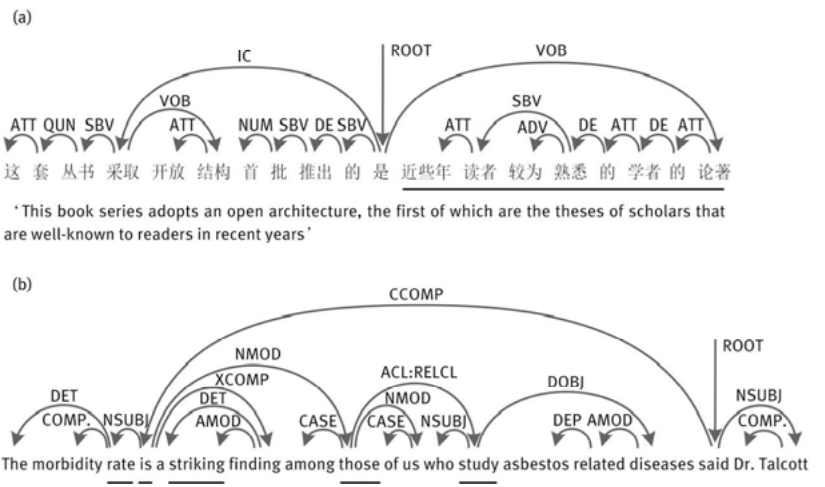


Fig. 7: (a) a unique Chinese sentence: $V[k] = 0.41$, MDD = 1.83 (SL = 19); (b) an English sentence: $V[k] = 1.04$, MDD = 2.50 (SL = 19).

3.4 The relationship between dynamic valency and MDD

Since $V[k]$ and $\langle k^2 \rangle$ in the random language are related to sentence length, according to the equation proposed by Ferrer-i-Cancho (2013 2014), $\langle d \rangle_{\min}$ (the minimum value of MDD) is only related to sentence length. However, the infer-

ence based on random languages may not apply to natural languages. Hence, the statistical analysis of the relationship between MDD and $V[k]$ in natural languages is necessary.

According to the equation of $\langle d \rangle_{\min}$ (Ferrer-i-Cancho 2013), we suppose MDD may increase with $\langle k^2 \rangle$ in natural languages. Meanwhile, $V[k]$ is related to $\langle k^2 \rangle$. Based on the scatterplot of the relationship between MDD and $V[k]$ (as shown in Fig. 8), we assume that the MDD is linearly correlated to $V[k]$, i.e. $MDD = aV[k] + b$. After removing the point with most fluctuation ($V[k] = 11.14681$) in Chinese, we performed regression analysis of the linear function between MDD and $V[k]$, with the results shown in Fig. 8.

We can observe that MDDs increase with $V[k]$, and the relationship between $V[k]$ and MDD in Chinese resembles that of English. By examining each $V[k]$ and its corresponding MDD, we find that almost every MDD in Chinese is larger than English. The regression equation of MDD is $y = 0.4337x + 2.2244$ ($R^2 = 0.467752$, P-value = 0.00) for Chinese, and $y = 0.3798x + 1.7882$ ($R^2 = 0.708659$, P-value = 0.00) for English. Here, the parameters a and b exhibit little difference in both languages. When the fluctuating tail of line is cut and $V[k] \leq 2.5$, the line of Chinese seems to go horizontally, while the regression equation of English matches the original data better, reflected by R^2 as well. Though the number of sentences in both treebanks varies, a finding can be obtained from Fig. 8 that the MDDs in both languages gradually increase with $V[k]$, displaying a linear relationship.

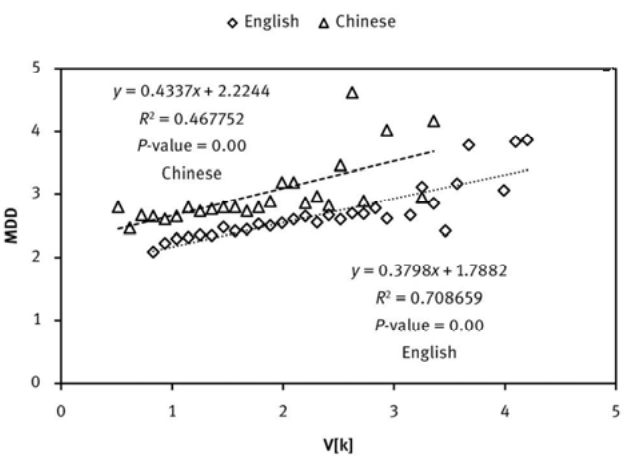


Fig. 8: Linear regression analysis of the relationship between MDD and $V[k]$ in the Chinese and English treebanks (SL = 19)

We have presented a preliminary explanation of the fact that Chinese has a larger MDD but a smaller $V[k]$. In fact, lower $V[k]$ means the absence of heads with greater out-degrees in the sentence. Given that there are more intervening words, the MDD will increase. Generally, higher $V[k]$ means there are more heads with higher degree in a sentence. If the dependents circle around their head or split the sentence into several chunks, the rise of $V[k]$ slows down in such a way to prevent the rapid rise of MDD, thereby producing a lower MDD of a language.

Therefore, though human languages tend to minimize the dependency distance, this optimization does not reach its limit in all human languages (Ferrer-i-Cancho 2006). The fundamental reason is that: language is a complex adaptive system as a result of the interaction of different factors in the cognitive mechanism and language use, which facilitates language production and understanding and results in particularities in human languages. In Chinese, a larger MDD partly results from some unique syntactic structures, including long modifiers preceding nominal constituent and a head on the tail of sentence. Such structures render more long-distance dependencies, that is, the connection between heads and dependents crosses over more intervening words, which leads to lower $V[k]$ but not necessarily lower MDD in Chinese. As our analysis illustrates, the sentence with lower $V[k]$ (indicating tall syntactic trees, namely, trees with greater hierarchical depths) does not always mean lower MDD, except for right or left branching.

In summary, the linear regression analysis shows that there is a positive correlation between the variance of dynamic valency and MDD. It remains to be further studied that whether the minimum $V[k]$ can successfully reduce the growth of MDD in Chinese. However, the mere fact that Chinese has a higher MDD may not suggest Chinese is more difficult to understand than English (Jiang & Liu 2015). As the frequently used grammatical regularities and patterns have been embedded in the long-term memory in the brain, the psychological predictability is formed during language comprehension (Liu et al. 2017). Similarly in English, despite the words with higher dynamic valencies, we may psychologically predict the valency of these words in linear order based on the existing prior knowledge. So a slightly longer dependency distance does not always cause difficulty in comprehension. Here, we can see that MDD is perhaps a metric to roughly estimate the overall difficulty of syntactic structure. Special cases call for more discussion. Concerning issues and inferences in this paper need further verification through psychological and cognitive experiments.

4 Conclusion

This paper examines the relationship between dynamic valency and DD in the Chinese and English dependency treebanks. Previous study reports that in random language there may be a positive correlation between the variance of degree and MDD (Ferrer-i-Cancho 2013). However, we find that the variance of dynamic valency ($V[k]$) in random dependency trees increases very slowly as sentence length rises. Our study based on natural language treebanks shows that the dynamic valency of words plays quite a different role in Chinese and English syntactic structures. By analyzing the relationship between MDD and $V[k]$ with various sentence lengths, we infer that Chinese has some unique syntactic structures absent in English, indicating that a smaller number of words with high dynamic valencies do not mean a lower MDD.

The questions proposed at the beginning of this paper have been answered:

(1) There is a difference in the dependency distance in Chinese and English at different sentence lengths, and dependency distance changes with sentence length, echoing the conclusion reached in the previous research.

(2) Different from the related literature, however, we discover that $V[k]$ at different sentence lengths in Chinese is no greater than that in English. It suggests that smaller $V[k]$ in natural languages does not mean smaller MDD, i.e., flatter trees do not mean higher MDD. There are special dependency structures in Chinese characterized by fewer words with high dynamic valency, namely smaller $V[k]$. Grammatical regularities, such as long modifiers preceding the head of the nominal constituent and the head placed at the tail of a sentence, result in more long-distance dependencies.

(3) Statistical analysis of the sentences at a specific length reveals that $V[k]$ has a positive correlation with MDD. To be more specific, the relationship between $V[k]$ and MDD in Chinese and English can be described in similar linear functions. However, the intercepts of the functions in both languages display a difference due to a number of shared $V[k]$ values and a smaller MDD of English than Chinese. Combined with the findings of (1) and (2), we perceive that various factors (syntactic structure, valency, etc.) other than dependence distance shall be taken into account when it comes to judging syntactic comprehension difficulty. It is supposed that the words with higher dynamic valencies do incur an increase in MDD. However, this effect on MDD may be suppressed

by other grammatical factors including word order, chunk length and, hierarchical depth, and so on.

This article throws light on the relationship between dynamic valency and dependency distance in human languages, especially universality and diversity in structure. Through comparative analysis, we try to reveal the laws of cognitive mechanisms and other constraints that affect the syntactic patterns.

There are some limitations that may render risk of errors in our analysis, including different annotation schemes, unequivalent semantic meaning in both languages, and different numbers of sentences in the news corpora (especially insufficient language material in Chinese). Admittedly, large-scale parallel treebanks which suit our study best are not available. Therefore, some conclusions of this paper remain to be supported when more suitable treebanks are available.

Acknowledgement: This work is supported by the National Social Science Foundation of China under Grant No. 17BYY120.

References

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

Ferrer-i-Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical Review E, 70* (5), 056135.

Ferrer-i-Cancho, Ramon. 2006. Why do syntactic links not cross? *Europhysics Letters, 76* (6), 1228–1235.

Ferrer-i-Cancho, Ramon. 2013. Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics, 25*, 1–21.

Ferrer-i-Cancho, Ramon. 2014. A stronger null hypothesis for crossing dependencies. *Europhysics Letters, 108* (5), 58003.

Ferrer-i-Cancho, Ramon & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5* (2), 143–155.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS, 112* (33), 10336–10341.

Gildea, Daniel & David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science, 34* (2), 286–310.

Hays, David. 1964. Dependency theory: A formalism and some observations. *Language, 40* (4), 511–525.

Herbst, Thomas. 2007. Valency complements or valency patterns? In Thomas Herbst & Katrin Götz-Votteler (Eds.), *Valency: Theoretical, Descriptive and Cognitive Issues* (pp. 15–36). Berlin/New York: de Gruyter.

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

Hudson, Richard. 1995. Measuring syntactic difficulty. Unpublished paper. Available from: http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf [2017 – 11 - 6].

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

Jiang, Jingyang & Haitao Liu. 2015. The Effects of sentence length on dependency distance, dependency direction and the implications. *Language Sciences, 50*, 93–104.

Kato, Akihiko, Shindo Hiroyuki & Matsumoto Yuji. 2016. Construction of an English dependency corpus incorporating compound function words. In *Proceedings of 10th edition of the Language Resources and Evaluation Conference* (pp. 1667–1671). Portorož: Slovenia.

Lecerf, Yves. 1960. Programme des conflits - modèle des conflits. Rapport CETIS, N4, Euratom, 1–24.

Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics, 15*, 1–12.

Liu, Haitao. 2008a. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9* (2), 159–191.

Liu, Haitao. 2008b. The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics & Its Applications, 387* (12), 3048–3058.

Liu, Haitao. 2009a. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.

Liu, Haitao. 2009b. Probability distribution of dependencies based on Chinese dependency treebank. *Journal of Quantitative Linguistics, 16* (3), 256–273.

Liu, Haitao. 2011. Quantitative properties of English verb Valency. *Journal of Quantitative Linguistics, 18* (3), 207–233.

Liu, Haitao. 2014. Language is more a human-driven system than a semiotic system. Comment on modeling language evolution: Examples and predictions. *Physics of Life Reviews, 11*, 309–310.

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

Lu, Qian, Chunshan Xu & Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages? *Complexity, 21* (s2), 33–41.

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

Miller, George A. 1956. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63* (2), 81–97.

Qiu, Likun, Yue Zhang, Peng Jin & Houfeng Wang. 2014. Multi-view Chinese Treebanking. In Jan Hajic & Junichi Tsujii (Eds.), *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014)* (pp. 257–268). Dublin: Centre for Global Intelligent Content.

Rosen, Kenneth H. 2012. *Discrete Mathematics and Its Applications* (7th Ed.). New York: McGraw-Hill.

Temperley, David. 2007. Minimization of dependency length in written English. *Cognition, 105* (2), 300–333.

Temperley, David. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics, 15* (3), 256–282.

Zhang, Hongxin & Haitao Liu. 2017. Motifs of generalized valencies. In Haitao Liu & Junying Liang (Eds.), *Motifs in Language and Text* (pp. 231–260). Berlin/Boston: de Gruyter .

Zipf, George. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Mass: Addison-Wesley Press.

Appendix

(a) For the five sampling sentences unique in the Chinese treebank, MDD = 2.79, $V[k]$ = 0.40997 (SL = 19). The sentences in the Chinese treebank are numbered after processing punctuation marks.

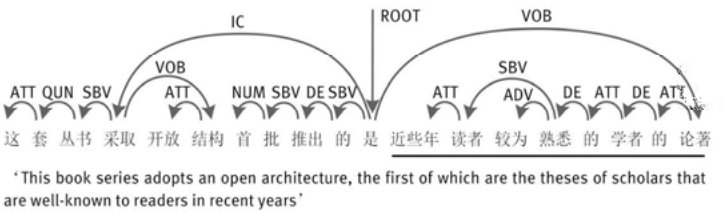


Fig. 9: Sentence 1, No.2053, MDD = 1.83, $V[k]$ = 0.40997 (SL = 19)

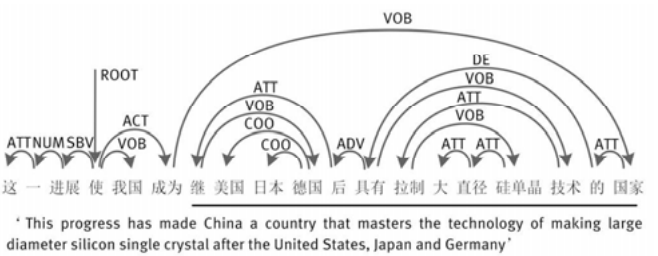


Fig. 10: Sentence 2, No.3651, MDD = 2.83, $V[k]$ = 0.40997 (SL =19)

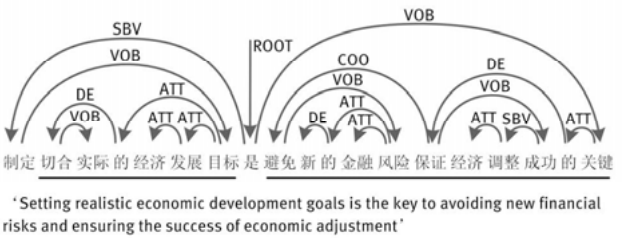


Fig. 11: Sentence 3, No.3967, MDD = 3.06, $V[k]$ = 0.40997 (SL = 19)

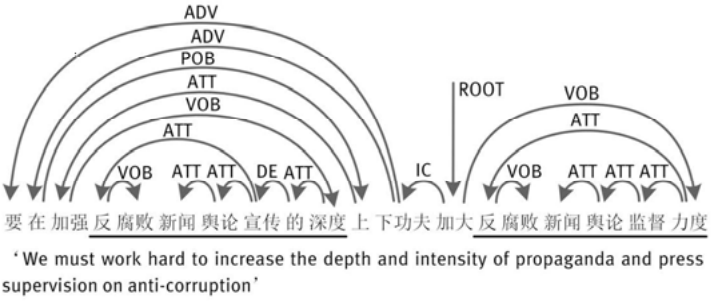


Fig. 12: Sentence 4, No.9493, MDD = 3.89, $V[k]$ = 0.40997 (SL = 19)

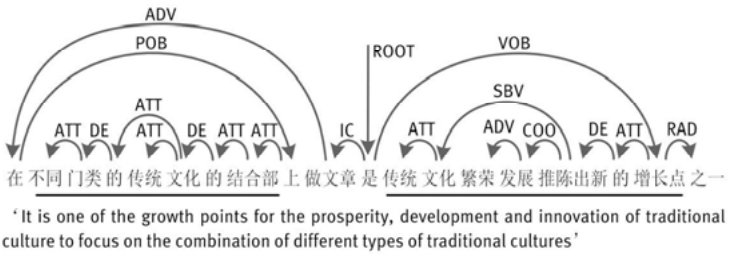


Fig. 13: Sentence 5, No.10597, MDD = 2.33, $V[k]$ = 0.40997 (SL = 19)

(b) As for the seven sampling sentences unique in the Chinese treebank numbered as 2432, 3003, 5018, 6692, 7361, 8276, 14199 respectively, $V[k]$ = 0.51524 and MDD = 2.79 (SL = 19).

(c) The 13 sampling sentences unique in the Chinese treebank are numbered 846, 1460, 1462, 2491, 3663, 7421, 8693, 8779, 9332, 9337, 11339, 13499 respectively. At the sentence length of 19, $V[k]$ = 0.62050 and MDD = 2.46.

Jingyang Jiang*, Jinghui Ouyang

Minimization and Probability Distribution of Dependency Distance in the Process of Second Language Acquisition

Abstract: Dependency distance minimization (DDM) is found as a universal quantitative property of natural languages. To investigate whether second language learners develop their interlanguage system under the pressure of DDM, we selected 367 Chinese EFL learners of nine consecutive grades, built one second language dependency treebank and two corresponding random treebanks and fitted different probability distribution models to dependency distances. It was found that: (1) The mean dependency distance (MDD) of interlanguage increases significantly across nine grades and the MDD of high-level learners doesn't reach the level of English native speakers. (2) The MDDs of interlanguage at different learning phases are significantly lower than their corresponding random languages (RL1 and RL2), indicating that learners develop their English proficiency under the pressure of DDM. (3) The distribution of dependency distances of RL1 cannot fit the Zipf-Alekseev distribution, but that of RL2 can. The parameters in the Zipf-Alekseev distribution of RL2 have no correlation with learners' language proficiency.

Keywords: dependency distance minimization; probability distribution; second language acquisition; writing

1 Introduction

Dependency distance refers to the linear distance between two linguistic units having a syntactic relationship within a sentence (Heringer et al. 1980; Hudson 1995). The linear distance between two words with syntactic relationship is thus restrained by human working memory. The results of corpus-based research and psychological experiments have indicated that human languages have a tendency towards dependency distance minimization (DDM) (Liu et al. 2017). This tendency suggests that, although human languages differ in pronuncia-

Jingyang Jiang, Zhejiang University, Hangzhou, P.R.China, jy-jiang@zju.edu.cn
Jinghui Ouyang, Zhejiang University, Hangzhou, P.R.China

https://doi.org/10.1515/9783110573565-009

tion, vocabulary, grammar, etc., their syntax may be constrained by universal mechanisms, and their evolution may have a universal model (Lu & Liu 2016a).

Dependency distance minimization is found as a universal quantitative property of more than 30 human languages (Liu 2008; Futrell et al. 2015). Moreover, text in all genres abides by the principle of DDM (Wang & Liu 2017). However, the materials that previous studies have investigated are all from native speakers' first language, and there is no such study on second language learners' language system. Compared with native languages, second language learners' language system owns its unique characteristics. As a dynamic language system, second language learners' language system, defined as 'interlanguage', is a structurally intermediate status between the native and target languages (Selinker 1972; Brown 1994). It can fossilize, or cease developing, in any of its developmental stages (Selinker 1972; Long 2008).

Human languages, as complex systems, seem to have evolved to come up with diverse syntactic patterns under the universal pressure for DDM (Liu et al. 2017). Along with the improvement of their second language proficiency, learners' second language gradually develops towards native speakers' language. Then, does second language learners' language system also develop under the universal pressure of DDM? If second language learners' language system gradually becomes more native like from primary learning stage to advanced learning stage, how will the development be reflected in the dependency distance?

To figure out how second language learners obey the principle of dependency distance minimization during the process of second language acquisition (SLA), we investigated the development of MDDs in Chinese EFL (English as a Foreign Language) learners' English writings at different learning phases to answer the first two specific research questions:

Question 1. How does the mean dependency distance in the English compositions written by Chinese EFL learners develop across nine grades?

Question 2. Does Chinese EFL learners develop their English proficiency under the pressure of dependency distance minimization?

Dependency distance can reflect the comprehension difficulty of syntactic structure (Liu 2008). DDM is considered as resulting from human cognitive mechanism (Liu 2008; Lu et al. 2016) and the effect of 'the principle of least effort' on syntactic structure (Zipf 1949). It also shows that humans prefer to avoid the use of long-distance dependencies to reduce cognitive cost. As a result, dependency distance distribution may present a certain pattern (Lu & Liu 2016a). Numerous previous studies (Liu 2008; Jiang & Liu 2015; Lu & Liu 2016a; Liu et al. 2017; Ouyang & Jiang 2017) have shown that the distribution of dependency distances presents certain regularity. Moreover, our previous study

(Ouyang & Jiang 2017) found that the probability distribution of the dependency distance of second language learners' interlanguage can well fit the Zipf-Alekseev distribution and the parameters a and b in the Zipf-Alekseev distribution well reflect second language learners' language proficiency at different learning stages. The Zipf-Alekseev function is verified to be the same model followed by the length distribution of many linguistic units (Popescu et al. 2014). To confirm that the previous findings (Ouyang & Jiang 2017) are not statistical artefacts and to further demonstrate the DDM of interlanguage in the process of SLA from the probability of dependency distances, we constructed two random languages using the compositions at different learning stages and fitted their probability of dependency distances to different exponential distribution and power law distribution models, including the Zipf-Alekseev distribution. We then try to solve other specific questions as follows:

Question 3. Does the probability distribution of dependency distance of random languages of second language learners' writings well fit the Zipf-Alekseev distribution? If the answer is yes, can the parameters in the Zipf-Alekseev distribution well reflect second language learners' language proficiency at different learning stages?

2 Materials and Methods

This study tries to answer the above-mentioned three research questions. Detailed information about the study's method and materials, including participants, materials, procedures and data analysis are described as follows.

2.1 Participants

Participants were 367 Chinese students from two high schools and one university in Zhejiang Province, China. Participants' basic information, including the number and the age of each grade is presented in Tab. 1. The subjects range from first grade of junior high school to first grade postgraduates of English major, which spans 9 grades. In China, most students started learning English from fourth grade in the elementary school. After the second year in universities, Chinese students usually do not need to take any English lessons. So choosing Chinese junior high school students, senior high school students, undergraduates of first and second grade (non-English majors) and postgraduates of English majors as participants can help us observe almost the whole

process of English learning. To discover whether Chinese EFL learners' compositions will present the same parameters in the probability distribution as those of English native speakers, we also compared the compositions of English native speakers with those of Chinese first grade postgraduates of English major, who, in the current study, are deemed as the representatives of high-level EFL learners in China.

Tab. 1: A Brief Profile of Participants

Group	Number	Age	Years of English Learning
First Grade of Junior High School (J1)	60	12-13	3-4
Second Grade of Junior High School (J2)	60	13-14	4-5
Third Grade of Junior High School (J2)	44	14-15	5-6
First Grade of Senior High School (S1)	44	15-16	6-7
Second Grade of Senior High School (S2)	39	16-17	7-8
Third Grade of Senior High School (S3)	41	17-18	8-9
First Grade of University (U1)	25	18-19	9-10
Second Grade of University (U2)	28	19-20	10-11
First Grade Postgraduate of English Major (P1)	26	22-23	13-14

2.2 Materials

Our self-built dependency treebank contains 367 English compositions written by the above-mentioned participants within the prescribed time limit in the class, with a total of 58583 words, and about 6500 words in each grade. The compositions collected are narratives. The topics of the compositions are basically about their own experiences, such as 'My Weekend', 'An Embarrassing Experience', 'An Unforgettable Experience', 'An Annoying Experience', etc. We controlled the genre and the subject matter by assigning compositions of similar topics to make a better longitudinal comparison. The contrastive dependency treebank of native English was extracted from Wall Street Journal (WSJ) Corpus. We selected linguistic data randomly from WSJ Corpus and built four subcorpora with about 6500 words of each corpus as contrastive dependency treebanks.

2.3 Procedure

After the students finished the implemented writing tasks, we collected the data and inputted them into the computer. All the 369 compositions were kept in a TXT format. Each composition was labeled with a unique code indicating students' school, grade, number and the topic. Students' compositions were faithfully keyboarded into the computer in exactly the same way as they were, including capitalization, punctuation, spelling and grammatical mistakes. The POS (Part-of-Speech) annotation and dependency relation tagging were automatically done by Stanford Parser 3.6.0, a tagging software developed by Stanford University (Marneffe & Manning 2008). To meet the requirements of our research, we modified some of the Stanford typed dependencies and established a new syntactic relation system. Moreover, we made an error tagging system and labeled L2 writing errors, including lexical errors and grammatical errors.

Although Stanford Parser can provide an effective version of annotation of all the raw data, there still exist quite a few mistakes because on the one hand, the accuracy of the program does not reach 100%; on the other hand, those participants with relatively low language proficiency will be very likely to make language mistakes, increasing the inaccuracy of the program. The preliminary annotation was done by Stanford Parser. After that, we did the manual check and modification. Moreover, applying the exactly same tagging systems to annotate the four contrastive corpora makes our research more accurate and scientific.

To investigate whether the second language learners develop their English proficiency ability under the pressure of DDM, we also constructed two random dependency treebanks for natural interlanguage of each grade. We used the two methods proposed in previous studies (Liu 2008) to generate the random treebanks.

2.4 Data Analysis

The concept of (dependency) distance is often used in the syntactic analysis framework with phrases or dependency relations as its basic constituents. The present paper uses the syntactic analysis framework of dependency grammar in which sentence structure is analyzed using the dependency relations between words in a sentence (Tesnière 1959; Hudson 2007, 2010; Nivre 2006; Liu 2009). A dependency relation has three core properties: binary, asymmetry, and labeledness.

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. Fig. 1 is a directed acyclic graph which shows a dependency analysis of the sentence 'He must have good ideas'.

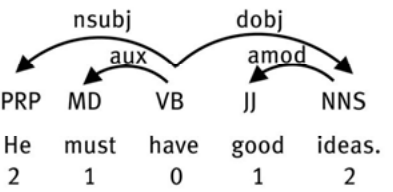


Fig. 1: Dependency structure of 'He must have good ideas.'

In Fig. 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 Fig. 1) depend on the nouns or verbs that they modify. 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.

The linear distance between governor and dependent is defined as 'dependency distance (DD)'. The concept was first used in (Heringer et al. 1980), who extracted the idea from the depth hypothesis of Yngve on phrase structure grammar (Yngve 1960, 1996). The term 'dependency distance' was introduced in Hudson (1995) and defined as 'the distance between words and their parents, measured in terms of intervening words'.

A method (Liu et al. 2009) was proposed 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 $W_1...W_i...Wn$ 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 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 DD the relevant measure is the absolute value of DD.

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

$$MDD(dependency\ type) = \frac{1}{n-1}\sum_{i=1}^{n-1}|DDi| \quad (1)$$

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.

For instance, a series of DDs can be obtained from the sentence in Fig. 1 as follows: 2 1 0 1 2. In other words, the example has two dependencies with DD = 2 and two dependencies with DD = 1. Using Formula (1), the MDD of this sentence is 6/4=1.5.

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

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

In this case, 'n' is the total number of words in the sample, and 's' is the total number of sentences in the sample. DD_i is the DD of the *i*-th syntactic link of the sample.

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 satisfactorily 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) provide 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. Fig. 2 shows a random analysis produced on this basis for the sentence in Fig. 1.

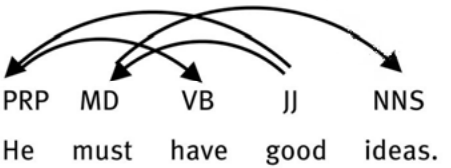


Fig. 2: A random analysis of *He must have good ideas* with crossing arcs

In the second random treebank (RL2), while governors are assigned randomly, we make sure that the resultant dependency tree (graph) 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). Fig. 3 gives an example of RL2.

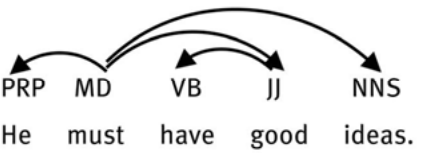


Fig. 3: A random analysis of *He must have good ideas* without crossing arcs

Thus, three dependency graphs can be constructed for the sentence 'He must have good ideas'. The first, shown in Fig. 1, is syntactically well-formed; RL1, shown in Fig. 2, has the lowest syntactic well-formedness, and RL2, in Fig. 3, exceeds RL1 in syntactic well-formedness.

In the previous studies of the distribution of dependency distances, the Zipf-Alekseev (Jiang & Liu 2015; Ouyang & Jiang 2017) was found well fitted the distribution model of dependency distances. The distribution of dependency distances is a kind of length distribution. Popescu et al. (2014) found that the Zipf-Alekseev function is adequate for the distribution of any linguistic units of physical length. Two assumptions were adopted by Hřebíček (1996, cited from Strauss and Altmann 2006):

The logarithm of the ratio of the probabilities P_1 and P_x is proportional to the logarithm of the class size, i.e.

$$\ln(P_1/P_x) \propto \ln x \tag{3}$$

The proportionality function is given by the logarithm of Menzerath's law (Hierarchy), i.e.,

$$\ln(P_1/P_x) = \ln(Ax^b)\ln x \tag{4}$$

yielding the solution

$$P_x = P_1 x^{-(a+b\ln x)}, x = 1,2,3,\ldots \tag{5}$$

If (3) is considered a probability distribution, then P_1 is the norming constant, otherwise it is estimated as the size of the first class, $x = 1$. Very often, diversification distributions display a diverging frequency in the first class while the rest of the distribution behave regularly. In these cases, one usually ascribes the first class a special value α, modifying (3) as

$$P_x = \begin{cases} \alpha, & x = 1 \\ \dfrac{(1-a)x^{(a+b\ln x)}}{T}, & x = 2,3,\ldots,(n) \end{cases} \tag{6}$$

where

$$T = \sum_{j=2}^{n} j^{-(a+b\ln j)}, a, b \in \Re, 0 < \alpha < 1.$$

Distributions (5) or (6) are called Zipf-Alekseev distributions. If n is finite, (6) is called a Right truncated modified Zipf-Alekseev distribution. In our project, we used the Altmann-Fitter software for fitting the model to the investigated data (Altmann-Fitter 2013).

3 Results and Discussions

3.1 Developmental Features of Dependency Distance at Different Grades

Tab. 2 shows the developmental MDDs across nine grades. The MDD of students' compositions at J1 is 1.841, J2 2.061, J3 2.064, S1 2.188, S2 2.219, S3 2.125, U1 2.422, U2 2.433 and P1 2.461. For a clearer illustration of the developmental MDDs, Fig. 4 is provided below. To explore whether the dependency distances increase significantly across nine grades, one-way ANOVA test and *post hoc* test

were employed. The results of one-way ANOVA test of dependency distances in L2 writings across nine groups of Chinese EFL learners indicate that the dependency distance in students' compositions increases significantly with the increase of the grade. Because MDD is not always increasing along with the increase of grades as shown in Fig. 4, *post hoc* test was conducted (see Tab. 2) in order to detect which adjacent groups have the significant difference.

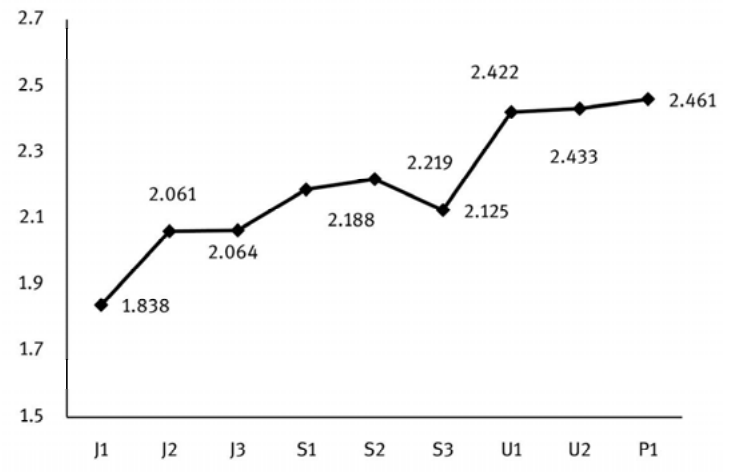


Fig. 4: Developmental tendency of MDDs at different grades

The statistical result showed that during the junior high school period, the MDD of Chinese EFL learners' English writings increases significantly (p=0.000) from J1 (1.841) to J2 (2.061), but stays stable (p=0.936>0.05) from J2 (2.061) to J3 (2.064). At senior high school, the MDD of Chinese EFL learner' English writings first increases significantly (p=0.003) at S1 (2.188), then continues increasing insignificantly (p=0.445>0.05) at S2, but experiences a significant (p=0.022) decrease at S3 (2.125). For university students, the MDD of their writings increases significantly (p=0.000) at first, but then keeps steady (p=0.782>0.005). There is no significant difference (p=0.476>0.05) between the MDD of the compositions by second grade non-English majors from that of postgraduate English majors with higher English proficiency.

Tab. 2: Descriptive statistics of dependency distances at different grades

Grade	Grade	Mean Difference	Sig.
(I)	(J)	(I-J)	
J1	J2	−0.220	0.000
J2	J3	−0.003	0.936
J3	S1	−0.124	0.003
S1	S2	−0.031	0.445
S2	S3	0.094	0.022
S3	U1	−0.297	0.000
U1	U2	−0.011	0.782
U2	P1	0.028	0.476

From the statistical analysis in Tab. 2, we can obtain the overall variations of the MDD of Chinese EFL learners' L2 writings across nine grades. The MDD of Chinese EFL learners' English writings increases significantly ($F(8, 50801) = 50.947$, $p=0.000$) from J1 to P1 as a whole, with a significant fall ($p=0.022<0.05$) at S3, but at the university level, the MDD stays stable with no significant progress.

Dependency distance, in the framework of dependency grammar, can be used as a metric of language comprehension difficulty (Liu 2008). MDD of a sentence is a good predictor of syntactic difficulty as found by the analysis of the dependency distance of sentences which present syntactic difficulty in psycholinguistic experiments (Liu 2008, Hudson 1995). In other words, in this case, the MDD can be used to reflect the syntactic complexity of students' English compositions. As a kind of natural language, theoretically, the syntactic complexity of second language learners' interlanguage can also be measured by dependency distance (Jiang & Ouyang 2017). The variations of the MDD indicate that the syntactic complexity of Chinese EFL learners' English writings increases along with the increase of their grades, or their English proficiency from J1 to P1, but decreases at S3. In the last year of the senior high school, as teachers and students focus on the review for college entrance examination, students don't have new language input and keep reviewing the learnt knowledge. The fall of MDD at S3 indicates that students' syntactic complexity decreases because of lacking new syntactic input. Students keep reviewing the same knowledge and there is no new comprehensible input that contains 'i+1'. According to Krashen's Input Hypothesis (2003), the learners progress in their knowledge of the

language when they comprehend language input that is slightly more advanced than their current level. Therefore, the fall of MDD for S3 students' compositions reflects that their English syntactic complexity doesn't improve.

Data analysis shows that the increase of MDD from U1 (2.422) to P1 (2.461) is very slight and insignificant (p=0.321), which indicates that the development of MDD in English writings of Chinese EFL learners comes to a steady state in university. From the junior grade, the MDD gradually increases, indicating a constant improvement of learners' syntactic capacity. However, when reaching a certain stage (university), the MDD doesn't increase and stays stable. This phenomenon can be explained by the Principle of Least Effort, according to which humans will avoid using dependencies of long distances that may cause more cognitive cost (Zipf 1949). These results indicate that when second language learners acquire a new language, they are also under the pressure of DDM. Second language learner's working memory capacity constrains their language comprehension and production during their process of second language acquisition (SLA) as native speakers do. This further illustrates Liu's et al. (2017) proposition that DDM is a human-driven linguistic universal resulting from cognitive mechanism and the effect of 'the principle of least effort' on syntactic structures.

To investigate whether the MDD of high-level Chinese EFL learners (postgraduate of English major) can reach the level of English native speakers, we compared the dependency distances in Chinese postgraduates' English writings with those in the contrastive four sub-corpora from WSJ and conducted independent T-test. The statistical results are presented in Tab. 3. The results of independent T-test show that there exist significant differences ($t_{(12490)}$= −1.426, p=0.002<0.05; $t_{(12471)}$= −1.089, p=0.017<0.05; $t_{(12223)}$=−3.047, p=0.000<0.01; $t_{(12302)}$= −1.628, p=0.000<0.01) between the dependency distances in English writings by Chinese postgraduates of English major (2.461±2.614) and those in four contrastive sub-corpora: WSJ1 (2.532±3.056), WSJ2 (2.516±2.952), WSJ3 (2.625±3.296)and WSJ4 (2.545±3.097).

The statistical results indicate that although the MDD of Chinese EFL learners' English writings has increased along with the improvement of leaners' English proficiency, the MDD of high-level Chinese EFL learners (postgraduate of English major) hasn't reached the level of English native speakers. Although Chinese postgraduate students of English major are considered as the high-level English learners in China, their syntactic complexity hasn't reached the level of English native speakers. This result agrees well with the findings of previous study (Lu & Ai 2015) that significant differences emerge between the native

speakers and college-level non-native speakers of different L1 backgrounds in all 14 syntactic complexity measures with the L2 Syntactic Complexity Analyzer (Lu 2010).

Tab. 3: The results of independent T-test between the dependency distances in English writings by Chinese postgraduates of English major and those in contrastive sub-corpora

	P1	WSJ1	WSJ2	WSJ3	WSJ4
Number of dependency distances	6198	6294	6274	6026	6105
Mean dependency distance	2.461	2.532	2.516	2.625	2.545
Standard deviation	2.614	3.056	2.952	3.296	2.545
F value of T-test		9.622	5.680	27.766	12.765
t value of T-test		0.002	0.017	0.000	0.000

The phenomenon that the MDD of high-level Chinese EFL learners stays stable but doesn't reach the level of native speakers illustrates that DDM is not the only reason accounting for the stagnation of learners' MDD. First, the stagnation of MDD may attribute to postgraduate English learners' interlanguage fossilization. Their English proficiency does not keep improving and enters a plateau period. In the case of interlanguage fossilization, their syntactic structural system does not seem to develop anymore. Moreover, postgraduates of English majors are not the best English learners in China. For example, those who get high scores in TOEFL or GRE tests may represent the highest-level English learners in China. And the MDD in their English compositions may be higher than that of postgraduates of English majors in this research.

3.2 The MDDs and the Distribution of Dependency Distances of Random Languages

Previous study (Liu 2008) assumed that the MDD and the mean number of item that are kept in working memory during the parsing are positively correlated. Liu associated MDD with working memory capacity, which has a value around 4 (Cowan 2001, 2005) and found that there is a threshold (less than 3 words) that the MDDs of most sentences or and texts of human languages doesn't exceed and it is within working memory capacity (Liu 2008). Considering second lan-

guage learners' interlanguage as a natural language, it seems reasonable to infer that second language learners' language system also has a minimized MDD threshold and it should also be within working memory capacity. If this is considered as a property of second language learners' language system, a random language should have a greater MDD than natural language. Therefore, we make a comparison between second language learners' language system and two random languages, and present the results in Fig. 5. which shows the MDDs of RL1 (crossing arcs) and RL2 (no-crossing arcs) of second language learners' interlanguage at each grade and of contrastive sub-corpora.

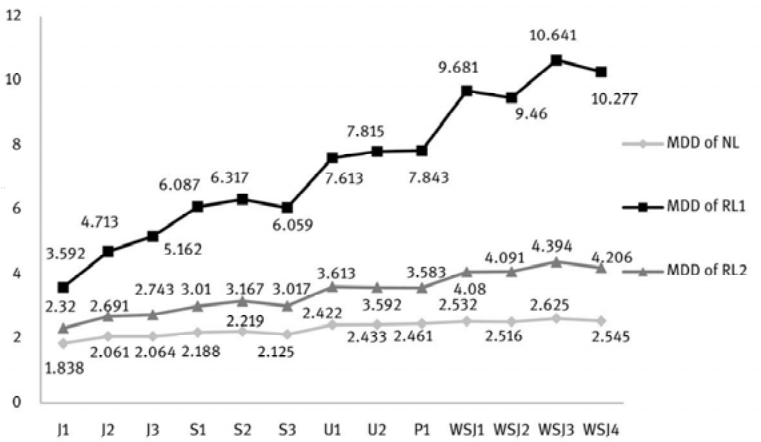


Fig. 5: MDD of interlanguage and random languages

Fig. 5 shows that for each language the two random analyses have much greater MDDs than their syntactic (NL) equivalents. Of the two random languages, RL2 has a smaller MDD than RL1, which agrees well with the results in previous study (Liu 2008) that 20 natural languages have smaller MDDs than those of their two random languages.

Apparently, at every grade level, the MDDs of second language learners' interlanguage and of RL2 are smaller than that of RL1. Ferrer-i-Cancho (2006) suggests that the uncommonness of crossings in the dependency graph could be a side effect of minimizing the Euclidean distance between syntactically related words. Meanwhile, the lower dependency distance can restrain the proportion of crossing dependencies (Lu & Liu 2006b). Our finding confirms that the projectivity can restrict the dependency distances (Ferrer-i-Cancho 2006). In

addition, the interlanguage at every grade level has a smaller MDD than RL2. This suggests that syntax also plays a critical role in minimizing the MDD of interlanguage. This indicates that at different language learning phases, second language learners obey the principle of DDM, which helps answer Question 3 that Chinese EFL learners develop their English proficiency under the pressure of DDM.

Our previous study (Ouyang & Jiang 2017) found that the Right truncated modified Zipf-Alekseev distribution well captures the probability distribution of dependency distance of Chinese EFL learners (the participants are the same in two studies) and the parameters a and b in the Zipf-Alekseev distribution well reflect second language learners' language proficiency at different learning stages, which also indicates that second language learners' syntactic acquisition process is always constrained by the tendency of dependency distance minimization. To confirm our previous findings (Ouyang & Jiang 2017) are not statistical artefacts and further demonstrate the DDM of interlanguage in the process of SLA from the probability of dependency distances, we constructed two random languages of interlanguage at different learning stages and fitted their probability of dependency distances to different exponential distribution and power law distribution models, including the Zipf-Alekseev distribution. If the random languages can well fit the Zipf-Alekseev distribution, will the parameters be related to second language learners' language proficiency?

Fig. 6 and Fig. 7 respectively show the distribution of dependency distances of RL1 and RL2 of nine grades and four sub-corpora from WSJ. Numerous previous studies have corroborated that the distribution of dependency distances of natural languages can well fit exponential distribution or power law distribution (Ferrer-i-Cancho 2004; Liu 2007; Jiang & Liu 2015; Lu and Liu 2016a; Ouyang and Jiang 2017). We fitted different exponential distribution and power law distribution models to the dependency distances of RL1 and RL2 of each grades and of WSJ. The quantities of different dependency distances of RL1 and RL2 were computed with quantitative linguistic software of Altmann-Fitter to determine the probability distribution models suitable for dependency distances of RL1 and RL2. The only problem is the chi-square goodness-of-fit test, whose reliability is increasingly doubted by linguists (Wang 2013; Mačutek 2013). It is inadequate if the sample size is very large, but no criterion as to what 'large' means has been given so far. We opt for the coefficient of variation C. If $C<0.02$, the fitting result is good; if $C<0.01$, the fitting result is very good (Liu 2017). It can be seen in Fig. 6 that the thirteen distribution curves are all concave down. But the fitting results of RL1 show that the dependency distances of thirteen groups of RL1 cannot well fit the same probability distribution.

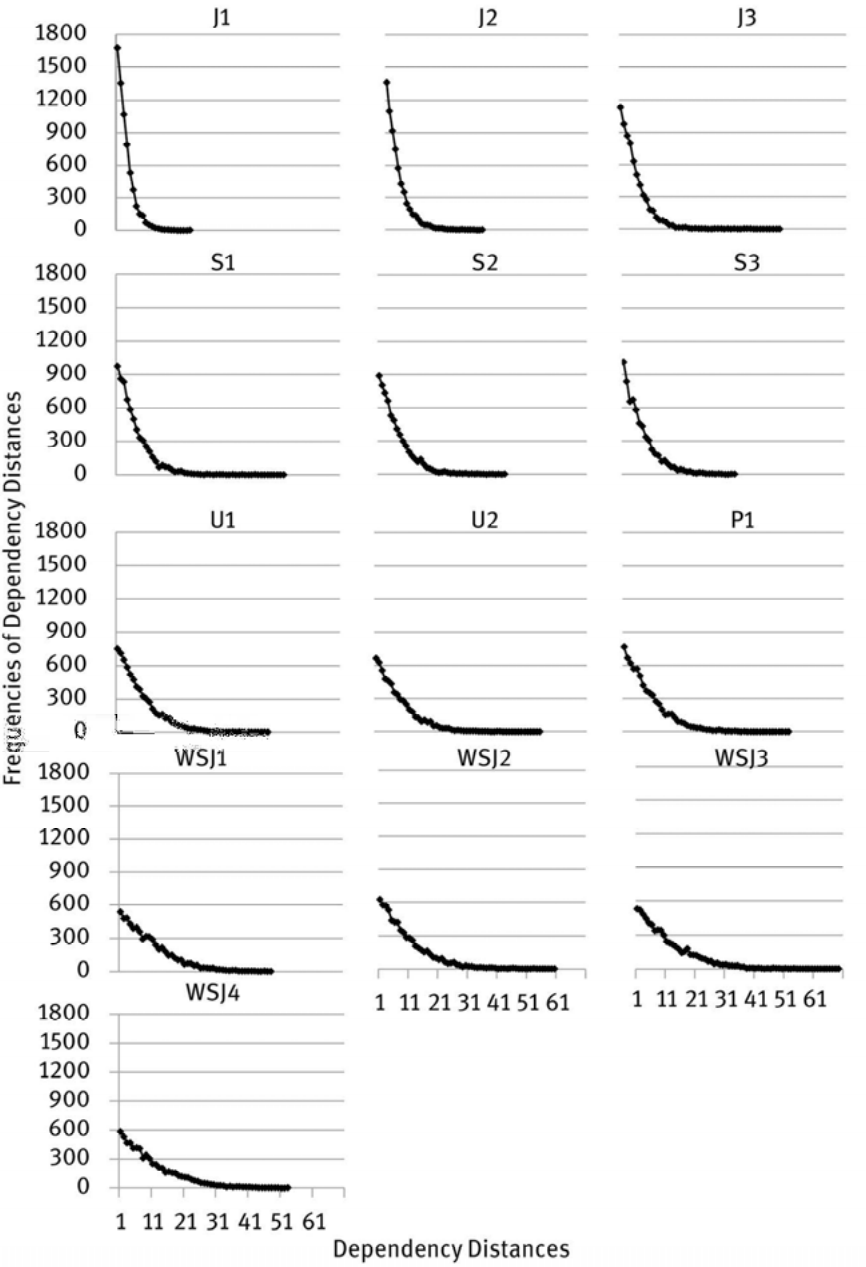


Fig. 6: The distribution of dependency distances of RL1

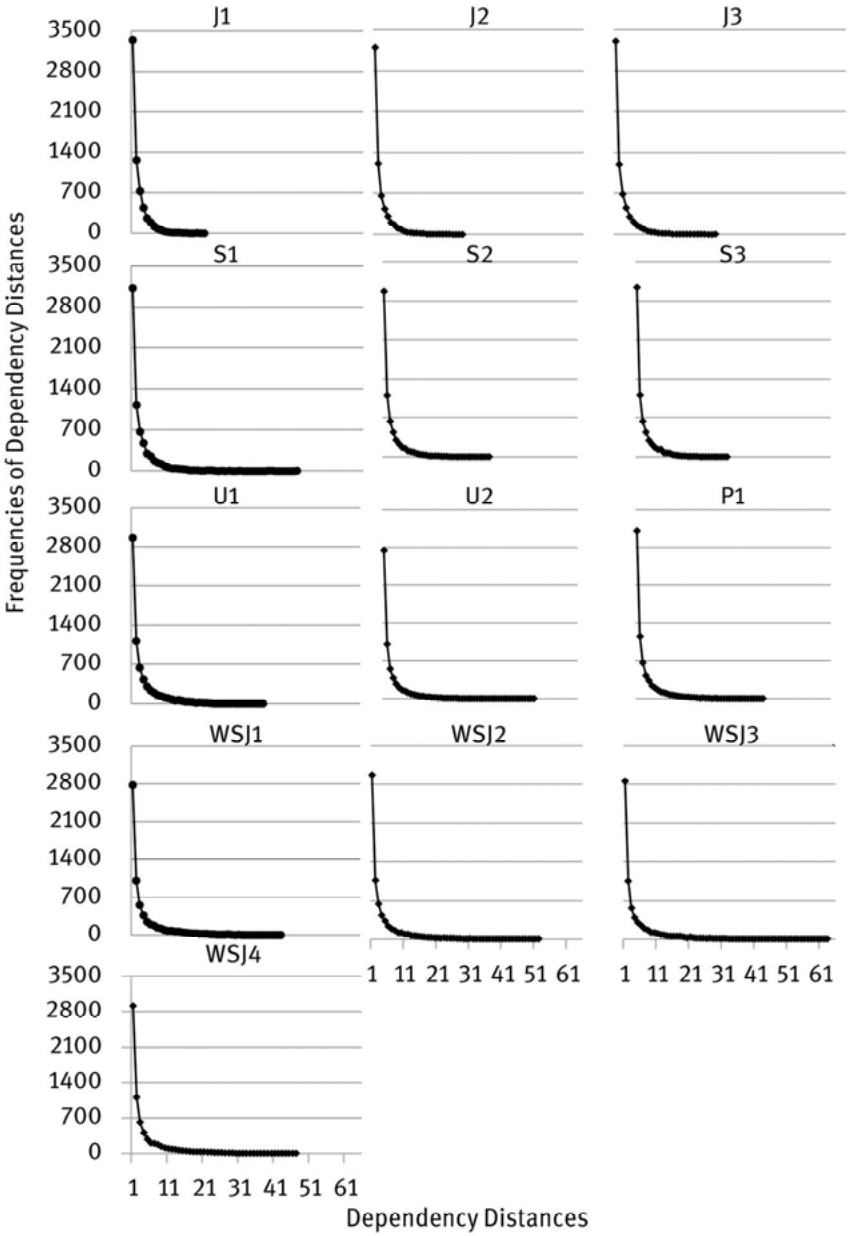


Fig. 7: The distribution of dependency distances of RL2

Likewise, the thirteen distribution curves of RL2 are all concave down. The fitting results show that the dependency distances of thirteen groups of RL2 can fit the following probability distributions: Right truncated modified Zipf-Alekseev (a, b; n=x-max, α fixed), Negative binomial (k, p), Right truncated negative binomial (k, p; R=x-max), Mixed negative binomial (k, $p1$, $p2$, α), Inverse Polya (a, k, p), Extended positive negative binomial (k, p; α fixed), Mixed geometric ($q1$, $q2$, α), and Mixed geometric-logarithmic (q, β, α). The mean values of R^2 of the thirteen groups with suitable models were calculated and tabulated as shown in Tab. 4. It can be seen that the probability distribution of dependency distances of RL2 can well fit the Right truncated modified Zipf-Alekseev.

Tab. 4: The determination R^2 of fitting of different models to the dependency distances of RL2

Groups of RL2	Right truncated modified Zipf-Alekseev	Negative binomial	Right truncated negative binomial	Mixed negative binomial	Inverse polya	Extended positive negative binomial	Mixed geometric	Mixed geometric-logarithmic
J1	0.9972	0.9999	0.9999	0.9999	0.9997	0.9999	0.9973	0.9999
J2	0.9991	0.9997	0.9997	0.9998	0.9996	0.9999	0.9994	0.9998
J3	0.9978	0.9998	0.9997	1.0000	0.9996	1.0000	0.9995	0.9998
S1	0.9953	0.9998	0.9998	0.9998	0.9991	0.9998	0.9989	0.9997
S2	0.9986	0.9998	0.9998	0.9998	0.9996	0.9998	0.9996	0.9998
S3	0.9986	0.9997	0.9997	0.9997	0.9992	0.9997	0.9996	0.9998
U1	0.9994	0.9993	0.9991	0.9991	0.9627	0.9996	0.9996	0.9995
U2	0.9995	0.9992	0.9990	0.9991	0.9702	0.9997	0.9993	0.9998
P1	0.9993	0.9992	0.9991	0.9992	0.9643	0.9999	0.9992	0.9997
WSJ1	0.9995	0.9981	0.9978	0.9979	0.9547	0.9997	0.9991	0.9996
WSJ2	0.9976	0.9991	0.9982	0.9989	0.9562	0.9998	0.9969	0.9992
WSJ3	0.9991	0.9981	0.9980	0.9982	0.9528	0.9993	0.9987	0.9998
WSJ4	0.9993	0.9978	0.9976	0.9976	0.9523	0.9996	0.9990	0.9994

In the previous study (Ouyang & Jiang 2017), it is found that the Right truncated modified Zipf-Alekseev distribution well captures the probability distribution of dependency distance of second language learners' natural languages at each grade and of native speakers' natural languages. Moreover, the parameters in the Right truncated modified Zipf-Alekseev distribution can well reflect second language learners' language proficiency at different learning stages. There are

altogether three parameters in the Right truncated modified Zipf-Alekseev distribution: a, b and α. The statistic results of the parameters are presented in Tab. 5. To obtain more accurate statistics of WSJ, we figured out the mean values of parameters of the four random sub-corpora of WSJ: a=1.0577; b=0.15105; α=0.4257.

Tab. 5: Fitting the Right truncated modified Zipf-Alekseev to the dependency distances of RL2

Groups of RL2	Parameters			X^2	$P(X^2)$	DF	N	C	R^2
	a	b	α						
J1	1.4607	0.2337	0.2555	112.3073	0.0000	16	21	0.0171	0.9972
J2	1.0562	0.2883	0.4833	73.5962	0.0000	23	28	0.0111	0.9991
J3	1.4234	0.1635	0.4867	116.5065	0.0000	24	29	0.0171	0.9978
S1	1.6004	0.0774	0.4626	288.9861	0.0000	42	47	0.0427	0.9953
S2	1.2210	0.1647	0.4515	135.4195	0.0000	31	36	0.0206	0.9986
S3	1.2507	0.1610	0.4627	144.5751	0.0000	26	31	0.0220	0.9986
U1	1.0434	0.1776	0.4327	133.6942	0.0000	33	38	0.0195	0.9994
U2	0.9620	0.2270	0.4520	95.0371	0.0000	46	51	0.0156	0.9995
P1	1.1694	0.1525	0.4405	111.4474	0.0000	38	43	0.0158	0.9993
WSJ1	0.9990	0.1740	0.4377	82.2867	0.0000	38	43	0.0130	0.9995
WSJ2	1.4866	0.0428	0.4261	167.3760	0.0000	47	52	0.0241	0.9976
WSJ3	0.8318	0.2012	0.4186	129.4442	0.0000	58	63	0.0190	0.9991
WSJ4	0.9134	0.1862	0.4204	101.5624	0.0000	42	47	0.0147	0.9993

A clearer illustration of the variations of parameters (a, b, α) is provided in Fig. 8. It is clearly shown that b and α remain relatively stable along with the grades, and a seems to decrease with the increase of the grade with several big fluctuations. To explore whether there is a correlation between Chinese EFL learners' English proficiency and the parameters a, b and α of RL2, we further did the correlation analysis between the grades and the parameters: a, b and α. The results of the data analysis show that there is no correlation between the parameter a and the grades (R^2=0.350, p>0.05), no correlation between the parameter b and the grades (R^2=0.119, p>0.05) and no correlation between the parameter α and the grades (R^2=0.074, p>0.05). Therefore, although the distribution of dependency distances of RL2 can well fit the Right truncated modified Zipf-Alekseev distribution, the parameters have no correlation with the grades.

This indicates that the parameters of the probability distribution of dependency distances of RL2 cannot measure second language learners' language proficiency; although it is found that the parameters of second language learners' natural languages can measure their language proficiency in our previous study (Ouyang & Jiang 2017).

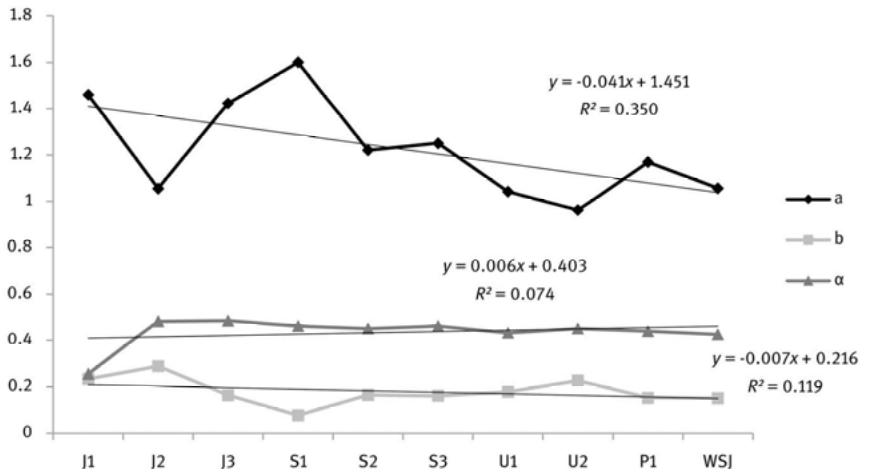


Fig. 8: The variations of parameters (a, b, α) of the Right truncated modified Zipf-Alekseev fitting the dependency distances of RL2

The fitting results of the dependency distances of RL1 and RL2 to the Zipf-Alekseev distribution further illustrates that the DDM is the principle only followed by second language learners' natural languages, but not their random languages.

It is interesting to note that the results of RL1 cannot fit the Zipf-Alekseev distribution, but the results of RL2 have good agreement with the Zipf-Alekseev distribution as natural languages. This agrees with the previous findings by Liu (2007) that the results of natural languages and their RL2 of six Chinese texts satisfy the Zeta distribution, but not RL1. Liu (2007) holds that projectivity is the background mechanism that causes this phenomenon. In Ferrer-i-Cancho's (2006) study, it is shown that projectivity can restrict the dependency distances. Obviously, in our study, projectivity is the background mechanism that makes the distribution of dependency distances of natural language and RL2 both be well captured by the Zipf-Alekseev distribution. RL2 has a lower MDD than RL1,

and natural language has a lower MDD than RL2. This suggests that syntax plays a key role in minimizing the MDD of second language learners' interlanguage system (Liu 2007). As compared with natural language, RL2 doesn't obey syntactic rules, though RL2 is projective as natural languages. We believe that the phenomenon that the parameters of natural languages can reflect second language learners' language proficiency, while the parameters of RL2 cannot can also be explained by syntax. Different from RL2 created by machine, second language learners' interlanguage is produced by humans. Syntax distinguishes the natural language from the artificial language. However, how syntax plays the role in the variations of the parameters of the probability distribution of dependency distance needs further research in the future.

4 Conclusions and Implications

Our data, derived from the cross-sectional dependency treebank of Chinese EFL learners' English writings from nine grades, suggest that the mean dependency distance (MDD) of Chinese EFL learners' English writings increases significantly across nine grades. The MDD increases from first grade of junior high school to postgraduate period as a whole, with a significant fall at third grade of senior high school, but stays stable at the university level, with no significant progress. The variations of the MDD indicate that the syntactic complexity of Chinese EFL learners' English writings progresses along with the increase of their grades, or their English proficiency from first grade of junior high school to first grade postgraduate of English major.

However, there exists a significant difference between the dependency distances of English writings by Chinese postgraduates of English major and those of news written by English native speakers, demonstrating that the MDD of high-level Chinese EFL learners (postgraduate of English major) doesn't reach the level of English native speakers. What's more, it is found that the MDD of Chinese EFL learners' English writings remain stable at the university level. We hold the view that it is caused on one hand by Chinese EFL learners' interlanguage fossilization and on the other hand the limit of working memory load, which causes dependency distance minimization (DDM) in this case.

Furthermore, the MDDs of learners' interlanguage at different learning phases are significantly lower than their corresponding random languages (RL1 and RL2). This indicates that Chinese EFL learners develop their English proficiency under the pressure of DDM. Furthermore, RL2 has a lower MDD than RL1, and natural language has a lower MDD than RL2, which suggests that syntax

also plays a key role in minimizing the MDD of second language learners' interlanguage system.

Moreover, the distribution of dependency distances of RL1 of second language learners' writings cannot fit any exponential distribution or power law distribution models. However, the distribution of dependency distances of RL2 and natural language can well fit the Zipf-Alekseev distribution. Projectivity is the background mechanism that causes this phenomenon. Besides, the parameters in the Zipf-Alekseev distribution of RL2 have no correlation with second language learners' language proficiency. This can be explained by syntax. Compared with natural language, though RL2 is projective as natural languages, it doesn't obey syntactic rules.

The current study corroborates that DDM is a language universal not only present in the use of first language, but also in the use of second language. This helps clarify the relationship between human cognition and second language. There is also a threshold that the MDDs of second language don't exceed and it is within working memory capacity. Studies on the dependency distances in relation to the cognitive demands on human cognitive system will remain the research focus for linguists or scholars in the field of cognition and psycholinguistics.

Acknowledgement: This work is supported by the National Social Science Foundation of China (17AYY021) and the Fundamental Research Funds for the Central Universities.

References

Altmann-Fitter. 2013. Altmann-Fitter User Guide. The third version. Downloadable at http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf (2016-8-29)

Brown, H. Douglas. 1994. *Principles of Language Learning and Teaching* (3rd Ed.). New Jersey: Prentice Hall.

Cowan, Nelson. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–185.

Cowan, Nelson. 2005. *Working Memory Capacity*. New York: Psychology Press.

Ferrer-i-Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical Review E, 70*(5), 056135.

Ferrer-i-Cancho, Ramon. 2006. Why do syntactic links not cross? *EPL*, *76*(6), 1228–1234.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS, 112*(33), 10336–10341.

Hays, David G. 1964. Dependency theory: A formalism and some observations. *Language, 40*(4), 511–525.
Heringer, Hans-Jürgen, Bruno Strecker & Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. Munich: Wilhelm Fink Verlag.
Hřebíček, Luďk. 1996. Word associations and text. In P. Schmidt (Ed.), *Glottometrika 15*(pp. 12–17). Trier: Wissenschaftlicher Verlag Trier.
Hudson, Richard. 1995. Measuring Syntactic Difficulty. Unpublished paper. Retrieved Oct. 4, 2016 from http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf
Hudson, Richard. 2007. *Language Networks: the New Word Grammar*. Oxford: Oxford University Press.
Hudson, Richard. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.
Jiang, Jingyang & 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.
Jiang, Jingyang & Jinghui Ouyang. 2017. Dependency distance: A new perspective on the syntactic development in second language acquisition. *Physics of Life Reviews, 21*, 209–210.
Krashen, Stephen D. 2003. *Explorations in Language Acquisition and Use*. Portsmouth: NH: Heinemann.
Lecerf, Yves. 1960. Programme des conflits - modèle des conflits. *Rapport CETIS. No. 4, Euratom*, 1–24.
Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics, 15*, 1–12.
Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.
Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.
Liu, Haitao. 2017. *An Introduction to Quantitative Linguistics*. Beijing: The Commercial Press.
Liu, Haitao & Fengguo Hu. 2008. What role does syntax play in a language network? *Europhysics Letters,* 83, 18002.
Liu, Haitao, Richard Hudson & Zhiwei Feng. 2009a. Using a Chinese treebank to measure dependency distance. *Corpus Linguistics and Linguistic Theory, 5*(2), 161–174.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews, 21*, 171–193.
Long, Michael H. 2008. Stabilization and fossilization in interlanguage development. In Catherine J. Doughty & Michael H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 487–535). Oxford: Blackwell.
Lu, Qian, Chunshan Xu & Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages? *Complexity, 21*(S2), 33–41.
Lu, Qian & Haitao Liu. 2016a. Does dependency distance distribute regularly? *Journal of Zhejiang University (Humanities and Social Sciences),* (4), 1–14.
Lu, Qian & Haitao Liu. 2016b. A quantitative study of the relationship between crossing and distance of human language. *Journal of Shanxi University (Philosophy & Social Science), 39*(4), 49–56.
Lu, Xiaofei. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496.
Lu, Xiaofei & Haiyang Ai. 2015. Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27.

Marneffe, Marie-Catherine & Christopher Manning. 2008. Stanford typed dependencies manual. Retrieved Oct. 7, 2015 from http://nlp.stanford.edu/software/dependencies_manual.pdf

Mačutek, Ján & Gejza Wimmer. 2013. Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics, 20*(3), 227–240.

Nivre, Joakim. 2006. *Inductive Dependency Parsing*. Dordrecht: Springer.

Ouyang, Jinghui & Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language probability of second language learners? *Journal of Quantitative Linguistics*. DOI: 10.1080/09296174.2017.1373991

Popescu, Ioan-Iovitz, Karl-Heinz Best & Gabriel Altmann. 2014. *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching, 10*(3), 219–231.

Strauss, Udo & Gabriel Altmann. 2006. Diversification – Laws in Quantitative Linguistics. Retrieved March 12, 2007, from http://www.uni-trier.de/uni/fb2/ldv/lql_wiki/index.php/Diversification

Tesnière, Lucien. 1959. *Eléments de la Syntaxe Structurale*. Paris: Klincksieck.

Wang, Lu. 2013. Word length in Chinese. In R. Köhler & Gabriel Altmann (Eds.), *Issues in Quantitative Linguistics* 3 (pp.39–53). Lüdendscheid: Ram Verlag.

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

Yngve, Victor. 1960. A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104*(5), 444–466.

Yngve, Victor. 1996. *From Grammar to Science: New Foundations for General Linguistics*. Amsterdam: John Benjamins.

Zipf, Kingsley, George. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Boston: Addison-Wesley Press.

Jingqi Yan

Influences of Dependency Distance on the Syntactic Development of Deaf and Hard-of-hearing Students

Abstract: Through dependency grammar approaches, the present study investigates the syntactic development in deaf and hard-of-hearing (DHH) students and intends to find out the possible mechanism that dependency distance minimization (DDM) functions on early syntactic performance. The results show that: 1) the right-truncated modified Zipf-Alekseev distribution best captures the probability distribution of dependency distance and the parameters may indicate language proficiency to some extent; 2) there's a general tendency for mean dependency distance (MDD) in higher grades of DHH group to approach near-native level; 3) MDD remains constantly low in different sentence lengths and in clause lengths in early school years in DHH group, while clause length and MDD have a more remarkable linear correlation, especially in higher-grade students from DHH group and normal hearing group. The findings, taken together, suggest a maximal tendency for DDM in early learning stages and a gradual development in DHH students' syntactic complexity.

Keywords: deaf and hard-of-hearing; dependency distance; syntactic complexity; DDM

1 Introduction

Dependency distance is defined as the linear distance between a dependent word and its head in a sentence (Liu 2008; Hudson 2010). It is now commonly accepted as an effective indicator of human cognition and of syntactic difficulty (Liu et al. 2017).

For one thing, dependency distance has been considered as an external reflection of working memory constraints and can be regarded as a metric to predict the memory load in language comprehension and production (Gibson 2000; Liu 2008). According to the hypothesis of dependency distance minimization (DDM) (Liu et al. 2017), there is a general tendency for language users to

Jingqi Yan, Zhejiang University, Hangzhou, P.R.China, jqyan@zju.edu.cn

https://doi.org/10.1515/9783110573565-010

minimize the dependency distance within the threshold of limited working memory, which may lead to shorter mean dependency distance (MDD) in natural languages than in artificial random languages and frequent use of short dependencies and rare use of long ones. Such long-tailed distribution of dependency distance has been well captured by many distribution laws, such as right-truncated zeta distribution, right-truncated waring distribution, exponential distribution (Jiang & Liu 2015; Liu 2008; Wang & Liu 2017) and right-truncated modified Zipf-Alekseev distribution (Ouyang & Jiang 2017).

For another, dependency distance and MDD may predict the syntactic difficulty (Hudson 1995). The greater the distance is, the greater difficulty the sentence structure has (Gibson 2000). Therefore, for less-skilled language users, they may not be able to use long-distance structures and use short dependencies instead. For example, several studies by Ninio (2011, 2014) found that in early language acquisition, adjacent dependency relations are more preferably used by children. The findings provide some clues about the possible representation of DDM in early language learning, as it suggests that when one is still developing his or her cognitive capacity or just starting to learn a new language, his or her writing may present an original or a simple form of DDM which is characterized by the greater dominance of adjacent relations. In a recent study (Ouyang & Jiang 2017), the distributions of dependency distance for writings of second language learners from different learning stages and for native speakers were fitted to the right-truncated modified Zipf-Alekseev distribution, and the result obtained an overall good fitting in the distributions for the writings of all the second language learners and native speakers. What's more, they also found a close correlation between the learners' second language proficiency and the parameters of the dependency distance distribution. The long-tailed distribution corroborates the existence of DDM in learners' languages and also suggests certain variations for the distribution in the learning course.

Despite its overall effectiveness in reflecting human cognition and syntactic difficulty, MDD is found to be influenced by another important syntactic indicator, i.e. sentence length (Ferrer-i-Cancho & Liu 2014; Jiang & Liu 2015; Wang & Liu 2017). For example, Jiang and Liu (2015) found that some traits of dependency distance may be influenced by sentence length, based on sentences of various lengths ranging from 10 to 30 words from a parallel corpus of Chinese and English. Further, Lu and Liu (2016) examined over 30 languages and found that the distribution of dependency distance of short sentences and of long sentences fit better to different models respectively.

By far most dependency analyses were conducted based on mature languages, i.e., utterances produced by proficient native language users with ma-

ture cognitive capacity. We still need to obtain more evidence about the variations of the quantitative features of dependency distance in individuals whose cognitive capacity and syntactic knowledge are not fully developed. For this concern, deaf and hard-of-hearing (DHH) students can be taken as potential research subjects. This study tried to extend the dependency distance study to the learners' written language by the deaf in particular. By comparing the writings of DHH adolescents with those of their hearing peers, we would like to investigate the possible regularities, differences and some changes in their dependency distance distribution and performance. We are particularly interested in trying to know whether the tendency of DDM may exist in learners' languages, and if it exists, how may DDM be represented in the language learning course, as can possibly be reflected in DHH individuals.

Working memory and cognitive capacity continue to develop and reach maturity during adolescence (Luciana et al. 2005). Therefore, it is interesting to observe how dependency distance may change during the process of adolescents' cognitive maturation. Although deaf individuals do not seem to differ in their working memory capacity (Lyxell et al. 2003), their language proficiency still lags far behind hearing peers (Musselman & Szanto 1998; Luckner & Handley 2008). More than 95% of DHH children are born to hearing parents and they only start to acquire sign language and learn print language after they go to school (Mitchell & Karchmer 2004). As a result, these children may not have full early access to language in either modality. For hearing children, the natural acquisition of syntax is completed in a short period before they go to school (de Villiers & de Villiers 1979), which makes it hard for researchers to observe some features of language development. By contrast, we may find the deaf an interesting subject of research because their natural acquisition of syntax is delayed and somewhat developed in a slow-motion picture. The study of language acquisition of deaf individuals may shed light upon the acquisition of general syntactic knowledge during the early years of language learning in this case. Therefore, in DHH individuals' adolescence, their written language shall exhibit more developmental characteristics.

In the present study, firstly, we would like to investigate the distribution of dependency distance in the writings of DHH students from different grades by comparison with the writings of normal hearing (NH) students of senior 3rd grade. Based on the findings by Ouyang and Jiang (2017), we may predict that the distribution of dependency distance may also follow certain models regardless of varieties in syntactic complexity while the parameters of certain distribution may indicate students' language proficiency. Also, it is expected that variations of MDDs in the DHH students' writings from different grades will be found.

Secondly, we try to investigate the relationship among MDD, sentence length and clause length and their changes between the two groups across different grades. In the writings of DHH individuals, sentence length can also indicate syntactic complexity. Combined results have suggested that DHH individuals tend to avoid the production of complex sentences and use short and simple sentences instead (Musselman & Szanto 1998; Wolbers et al. 2012). They may have shorter mean length of utterances and sentences than the typical population do (Nippold et al. 2008, 2009). Therefore, the sentence length and its influence on the language performance indicated by MDD should be taken into consideration in our syntactic study. On the other hand, current dependency studies usually use periods, exclamation marks and question marks as sentence segmentation markers. These are not sufficient to classify sentence boundaries in Chinese as commas are also similarly used to signal a sentence with complete structures in the language. There are many run-on sentences which contain various events in Chinese (Xue & Yang 2011). In this sense, sentence segmentation based on the end-of-sentence punctuations may lead to a misunderstanding of the relationship between sentence length and MDD. For deaf individuals, the proper use of punctuations is even more challenging. Therefore, their classification of end-of-sentence punctuations and within-sentence punctuations is vague, obscuring the influence of sentence length on MDD and syntactic difficulty as well. Regarding these problems, we consider both sentence length and clause length as factors to influence syntactic difficulty and MDD. To discover the possible manifestation of DDM and to figure out the syntactic development in DHH students, three major research questions are listed:

1) What is the probability distribution of the dependency distance in the writings of DHH students? Can parameters in the distribution model and MDD reflect DHH students' language proficiency across different grades?
2) Does MDD differ with the increase of sentence length? Does the relationship between MDD and sentence length change across different grades in DHH group?
3) Does MDD differ with the increase of clause length? Does the relationship between MDD and clause length change across different grades in DHH group?

2 Method

The concept of dependency distance (DD) is embedded in the concept of a dependency relation. A dependency relation has the following core features: it is a binary relation between two linguistic units with a direction; it is asymmetrical, with one of the two units acting as governor and the other as dependent; it is labeled by a dependency type on top of the arc linking the two units (Tesnière 1959; Hudson 1990; Liu 2009). To elucidate the features, Fig. 1 presents a dependency structure of a sample sentence in English.

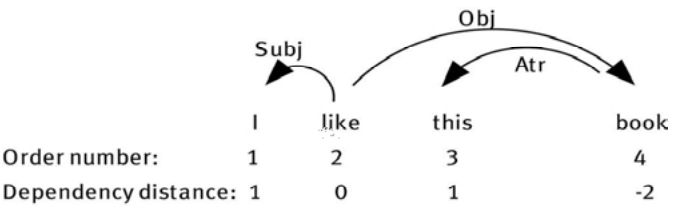


Fig. 1: The dependency structure of a sentence

In Fig.1, all the words are linked by labeled dependency types. The directed arc presents the relation between the two units. For example, the verb *like* governs the pronoun *I*, forming a subject (Subj) relation. *Like* therefore becomes the governor and *I* becomes the dependent. *Like* also governs the noun *book*, forming an object (Obj) relation. Then the adjective *this* is governed by *book* to form a relation of attribute (Atr).

Such binary dependency relation enables the measurement of DD. Formally, let $W_1 \ldots W_i \ldots W_n$ be a word string. When W_a and W_b forms a dependency relation, if W_a is a governor and W_b is its dependent, then the DD can be calculated as a–b. When 'a' is greater than 'b', the DD value is positive, which means that the governor follows the dependent; when 'a' is smaller than 'b', the DD value is a negative number and the governor precedes the dependent. For example, from Fig. 1, the second word (W_2) *like* governs the first word (W_1) *I*. Therefore, the dependency distance for the subject relation is 2–1=1. The positive and negative number is useful for distinguishing dependency direction. However, in measuring MDD the absolute value of DD is adopted. In addition, in a sentence, there is generally one word (the root verb) without a governor, whose DD is therefore defined as zero.

The mean dependency distance (MDD) of a collection of sentences can be defined as:

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

Here 'n' is the total number of words in the text, 's' is the total number of sentences in the sample. 'DD_i' is the dependency distance of the *i*th syntactic link of the text.

For effective measurement of DD for each dependency relation, in this paper, we built a dependency treebank based on the writing samples collected from DHH students (age 10–18, with profound hearing loss) from primary 4th to 6th grade , junior 1st to 3rd grade, and senior 1st to 3rd grade in a special education school. To compare their writings to those of the hearing students, writing samples were also collected from the senior 3rd hearing graders in a writing competition. The samples from DHH students are mainly students' compositions and diaries and those written by hearing students are mainly narrative compositions. With full annotation of part-of-speech, governors, dependent and dependency types guided by the Chinese dependency syntax in the treebank, the dependency distance for each dependency type can be calculated easily. To briefly illustrate the treebank building, the dependency structure in Fig. 1 can be converted to the format presented in Tab. 1.

Tab. 1: Annotation of a sample sentence in a treebank

Dependent			**Governor**			**Dependency type**	**DD**
Order number	Character	POS	Order number	Character	POS		
1	I	R	2	like	V	Subj	1
2	like	V	5	.	Bjd	S	
3	this	Adj	4	book	N	Atr	1
4	book	N	2	like	V	Obj	−2
5	.	Bjd					

Note. R = Pronoun, V = Verb, Adj = Adjective, N = Noun, Bjd = end-of-sentence punctuation marks, Subj = Subject, S = main sentence, Atr = Attribute, Obj = Object.

Each word token is vertically presented in the array marked by its linear order number in the sentence. Its governor word, also marked with its linear order, follows in the next column. The dependency type between the two words is annotated, and the corresponding dependency distance can hereby be obtained by subtracting the order number of the governor to that of the dependent. A positive dependency distance indicates a relation with the dependent preceding the governor, while a negative distance represents the opposite.

As the writing samples were learners' language, we also classified writing errors into 13 types (8 lexical error markers, 4 syntactic errors and 1 semantic ambiguity) in the dependency relation annotation. The final treebank contains 44421 words, 2854 sentences and 187 writing samples. The extraction of sentences is based on sentence segmentation by end-of-sentence punctuation, namely, the period, exclamation point, apostrophe and question mark. The extraction of clauses is based on clause segmentation by within-sentence punctuation, namely, comma and semicolon. Text information for each grade is presented in Tab. 2.

Tab. 2: Text information

Grade	Text Number	Word Count	Sentence Number
DP4	40	4339	283
DP5	20	2948	195
DP6	22	3860	250
DJ1	12	3408	167
DJ2	12	3488	253
DJ3	28	4969	416
DS1	13	3964	245
DS2	14	2707	169
DS3	11	5627	323
NH	15	9111	553

Note. DP4 = primary 4th grade in DHH group, DP5 = primary 5th grade in DHH group, DP6 = primary 6th grade in DHH group, DJ1 = junior 1st grade in DHH group, DJ2 = junior 2nd grade in DHH group, DJ3 = junior 3rd grade in DHH group, DS1 = senior 1st grade in DHH group, DS2 = senior 2nd grade in DHH group, DS3= senior 3rd grade in DHH group, NH = senior 3rd grade in NH group.

3 Results and discussion

3.1 Distribution of dependency distance

Some previous studies have suggested that even languages by DHH students and aphasic individuals may exhibit similar distribution patterns as native language users do (Jin & Liu 2016; Neophytou et al. 2017). We may firstly investigate the frequency distribution of dependency distance for all grades of DHH writings and of NH one using three different models that have been used and testified in other language samples.

The software Altmann-Fitter (2013) was adopted to test the distributions of dependency distance of all written sets. Three distribution models were fitted to the distributions of dependency distance.

Tab. 3 demonstrates the fitting results of determination coefficient R^2 for different written sets by right-truncated Zeta (a; R=x-max), right-truncated Waring (b, n) and right-truncated modified Zipf-Alekseev distribution (a, b; n=x-max, α fixed).

Tab. 3: R^2 of dependency distance distribution for each grade fitted by three models

	Right-truncated Zeta		Right-truncated Waring		Right-truncated modified Zipf-Alekseev	
	$P(X^2)$	R^2	$P(X^2)$	R^2	$P(X^2)$	R^2
DP4	0.0000	0.9960	0.0190	0.9991	0.4183	0.9997
DP5	0.0000	0.9977	0.1388	0.9993	0.2941	0.9997
DP6	0.0000	0.9971	0.0075	0.9993	0.0994	0.9993
DJ1	0.0000	0.9929	0.0522	0.9993	0.2527	0.9999
DJ2	0.0000	0.9957	0.0112	0.9986	0.1805	0.9998
DJ3	0.0000	0.9913	0.0993	0.9989	0.6777	0.9993
DS1	0.0000	0.9951	0.0001	0.9973	0.1396	0.9998
DS2	0.0000	0.9967	0.0002	0.9976	0.2750	0.9998
DS3	0.0000	0.9972	0.0000	0.9937	0.0679	0.9997
NH	0.0000	0.9972	0.0000	0.9978	0.1119	0.9998

The fitting results of dependency distance distributions are not excellent for all of the three models in all grades from both DHH and NH group as indicated by the $P(X^2)$ value, notwithstanding the overall high R^2. Only the right-truncated

modified Zipf-Alekseev distribution has the best fit, as the $P(X^2)$ value is larger than 0.05 for all the written sets of both DHH and NH group. The fitting may echo the previous finding that Zipf-Alekseev function is a good model for length distribution of many different linguistic units (Popescu et al. 2014). To capture a vivid picture of probability distribution of dependency distance, we present the fitting of right-truncated modified Zipf-Alekseev distribution of every written set in DHH and NH group (see Fig. 2). The fitting results show that despite the possible writing proficiency divergences across all grades between the two groups, their written language forms a system of excellent self-regulation and self-organization. The abundant use of adjacent relations, as shown by the distribution of dependency distance, supports the universal tendency of DDM.

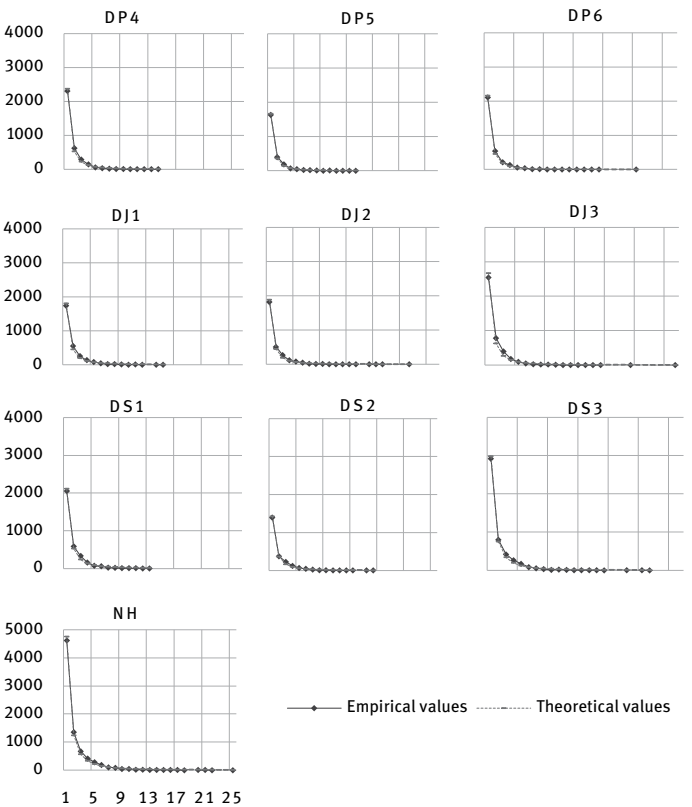


Fig. 2: Probability distribution of dependency distance in each grade from DHH and NH group. Numbers on the x-axis refer to the dependency distance, and those on the y-axis refer to the frequency of the DD.

If we inspect the distribution curve for each written set, individuals in lower grades do exhibit fewer uses of long dependencies, possibly leading to shorter MDDs. Some differences of different grades may be revealed by changes in the parameters. Therefore, to further explore the possible variations of parameters across different written sets, we conducted nonparametric Kruskal-Wallis tests to determine whether there are significant differences among grades presented by the three parameters of the model. Only a marginal significant difference among grades was found in parameter a of the right-truncated Zipf-Alekseev distribution, $X^2_{(9)} = 16.819$, $p = 0.052$. Seen from Tab. 4, mean value of α tends to be smallest in NH group and at senior grades in DHH group. The variations of parameters among grades were similarly found in the second language acquisition study by Ouyang and Jiang (2017). While Ouyang and Jiang found significant variations of parameter b across grades using a correlation analysis, such changes were only found in the parameter α in this study. This may suggest that for different languages, language proficiency may be indicated by different parameters for the Zipf-Alekseev model, though such implication may need more solid evidence by more future studies.

Tab. 4: Mean values of the three parameters of right-truncated modified Zipf-Alekseev Distribution

	Mean value of parameter a	Mean value of parameter b	Mean value of parameter α
DP4	0.681	0.278	0.596
DP5	0.852	0.309	0.628
DP6	0.760	0.166	0.609
DJ1	0.777	0.349	0.581
DJ2	0.838	0.304	0.595
DJ3	0.879	0.394	0.599
DS1	0.869	0.398	0.602
DS2	0.843	0.187	0.573
DS3	0.834	0.235	0.583
NH	0.765	0.282	0.569

One may further wonder, despite the parameter variations, how does MDD change across different grades? In the next sub-section, we would like to explore the MDD changes across grades in DHH students and in NH students.

3.2 MDD variations across different grades

To investigate the variations in the syntactic complexity among different grades, a one-way ANOVA was conducted with MDD as independent variable and grade as dependent between-group variable. Between-group differences were found ($F = 12.425$, $df1 = 9$, $df2 = 177$, $p < 0.001$). As can be seen from Fig. 3, there is a general tendency for MDD to increase as the grade rises, and the writings by the NH students have the longest MDD.

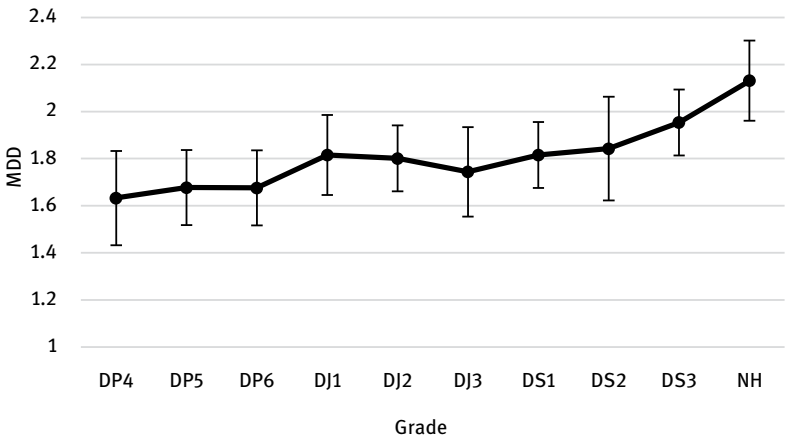


Fig. 3: MDD changes in each grade from DHH group and NH group

The MDD variations across grades support Liu, Xu and Liang's (2017) assumption that MDD can indicate syntactic complexity differences. NH students may use more complex structures with long dependency distances, while DHH individuals continue to struggle for the syntactic development. The writing of DHH students is short and simple, choppy, rigid, stereotyped, with fewer uses of complex sentence and structures (Wolbers et al. 2012). Such characteristics in their writing may lead to an overall short MDD. However, via frequent language practices, their syntactic complexity does show some improvement, as MDD value approaches NH level in the senior 3rd grade. Researches in mature languages have found certain relationship between sentence length and MDD (e.g. Jiang & Liu 2015; Lu & Liu 2016). But for DHH students whose languages are not proficient enough, can sentence length affect MDD in the same way in DHH students as in NH students? In the next sub-section, we will try to answer this question.

3.3 Relations between mean dependency distance, sentence length, and grades

Before we examine the inter-relation, we may firstly observe the proportion of sentences with different sentence lengths across all grades in DHH and NH group.

As can be seen from Fig. 4, the percentages of sentences, similar to Jiang and Liu's (2015) finding, exhibit rather inconstant variations for different sentence lengths. The majority of sentences have the length of no longer than 40 words.

The inconstant changes of sentence length in frequency use may suggest diversified use of sentences. However, the extremely long tail of sentence length distribution for all grades may not suggest greater uses of complex structures. After all, as suggested by the MDDs in the previous sub-section, syntactic complexity in the lower grades of DHH students is relatively lower than that of NH students. The abnormal presence of long sentences in DHH students' writings support our prediction as was noted in the Introduction section. DHH students may overuse run-on sentences and mistakenly use punctuations. This result may give rise to greater risks when examining the relationship between sentence length and mean dependency distance.

Based on the sentence length distribution, we chose sentence length from 3 to 40 words to examine the relationship between the MDD and sentence length in all the written sets, as about 93% of the sentences fall within this range of sentence length.

A linear model with sentence length as a predictor was first fitted to mean dependency distances. The model was highly significant with low statistical correlation ($F = 170.5$, $df1 = 1$, $df2 = 2512$, $p < 0.001$, adjusted $R^2 = 0.0632$). Then we added grade as a second predictor to form interaction with sentence length. The updated model was still highly significant with low correlation ($F = 24.37$, $df1 = 19$, $df2 = 2494$, $p < 0.001$, adjusted $R^2 = 0.1502$), though the correlation efficient increased to a certain degree. A likelihood ratio test was performed between the two models. The significant difference ($p < 0.001$) by comparison shows that the different grades from DHH and NH group, together with sentence length may have possibly influenced MDD. Fig. 5 presents the curve trend of MDD with the variations of sentence length in each grade from DHH and NH group.

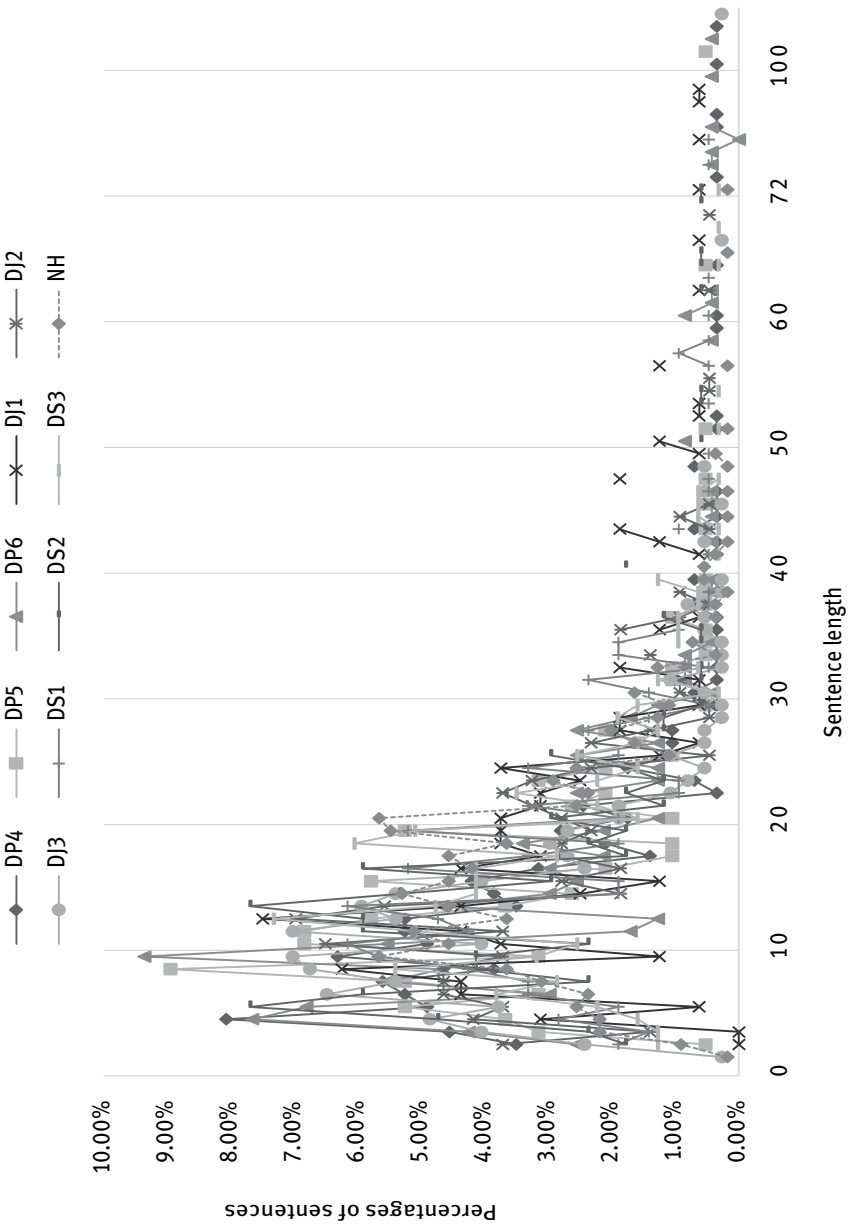


Fig. 4: Sentence length distribution

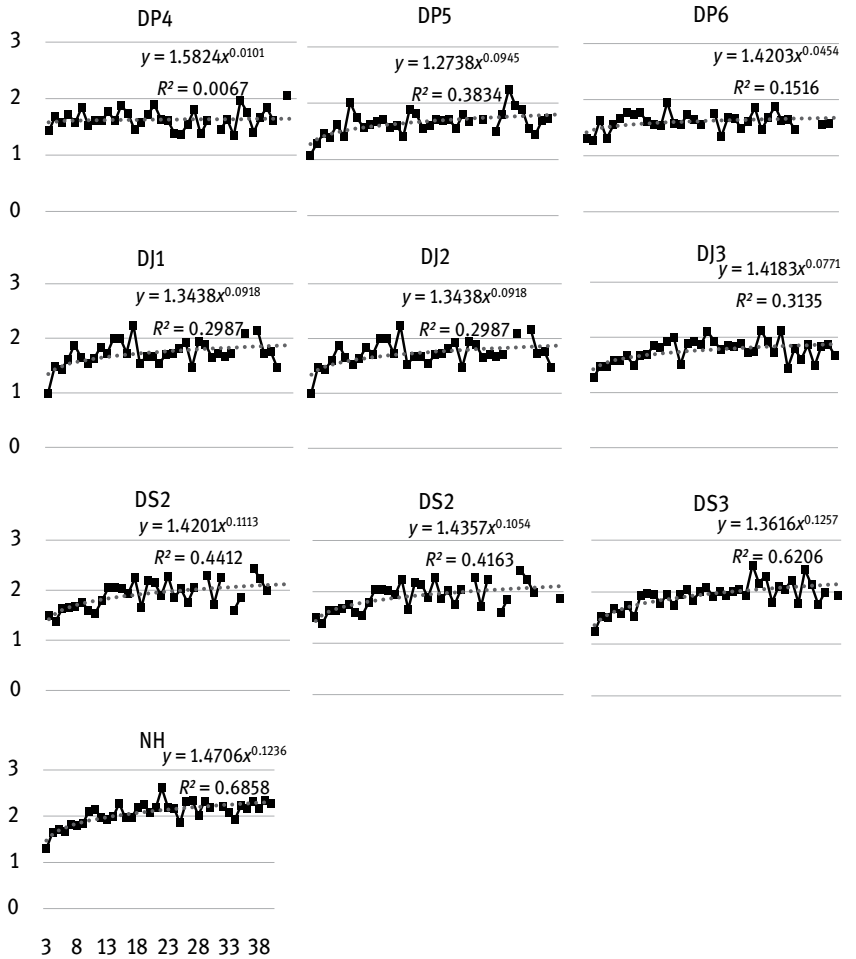


Fig. 5: MDD variations with different sentence lengths in each grade from DHH and NH group. Numbers on the x-axis refer to the sentence length, and those on the y-axis refer to the MDD.

From Fig. 5, the trend curves for the MDD changes with sentence length variation fit best to a non-linear slowing of the power function, which suggest that MDD has a steady and slow increase with the rise of sentence length. However, such tendency is not well observed in lower grades of DHH group, as they do not display a regular MDD change along with sentence length variations. The MDD, especially in the primary 4th grade, is constantly constrained within a narrow range even in the long sentences. Such situation gradually changes as

the grade increases. In the senior 3rd grade, the variation pattern becomes generally similar to that of NH group, as MDD slowly climbs up when the sentence length increases.

The result, to some extent, suggests that in the early learning stages, students tend to maximize their use of adjacent relations. DDM in the early stages seems to have the simplest manifestation due to the limited syntactic complexity and cognitive capacity (Yan 2017). No matter how long the sentence is, younger DHH students do not have the ability to produce more complex sentence structures with longer MDD. For senior 3rd grade students, they seem to be able to perform similarly with NH students, with comparable maturity in syntactic complexity. For these elder students, their MDD gradually increases when sentences are longer, suggesting a greater use of non-adjacent relations with more complex structures. These long sentences are produced by a compromise between the more complicated communicative need and the pressure of DDM (Liu et al. 2017).

3.4 Relations between mean dependency distance, clause length, and grades

The relationship between MDD and sentence length, as suggested in the previous sub-section, can sometimes be masked by run-on sentences and students' mistaken use of punctuations. Therefore, we would like to further examine the relationship between MDD and clause length across different grades.

Still, we may take a look at clause distribution first. As presented in Fig. 6, the clause length distribution distinguishes notably from sentence length distribution. The curves for clause length distribution are rather smooth, presenting a reversed U-shape distribution. For all grades, the highest proportion consistently sits at the clause length of 4. The largest use of clause length of 4 possibly echoes the central working memory storage limit of 4±1 unit by human beings (Cowan 2005, 2010). The 4-word clauses may be most comfortable and comprehensible for both the hearers and speakers. For younger DHH students, they feel more secure in using the 4-word structures. However, one must develop his or her complex use for more complicated expression of concepts.

To better examine the relationship between clause length and MDD, we chose clause length between 2 to 15 words, taking up the overall proportion of about 95%.

Similar to the treatment with sentence length, a linear model with clause length as independent variable and MDD as dependent variable was fitted. Again, the first linear model showed a significant *p*-value but a low correlation

(F = 677.6, $df1$= 1, $df2$=7618, p < 0.001, adjusted R^2 = 0.0816). A second regression followed adding grade as a new predictor. The model is significant as a whole, but correlation is quite low (F =52.56, $df1$= 19, $df2$= 7600, p <0.001, R^2 = 0.1139). By comparing the two models, likelihood ratio test suggests significant differences. The result indicates that grade and clause length may also be related with mean dependency distance. Fig. 7 presents the curve trend of MDD with the clause length variations in each grade from DHH and NH group.

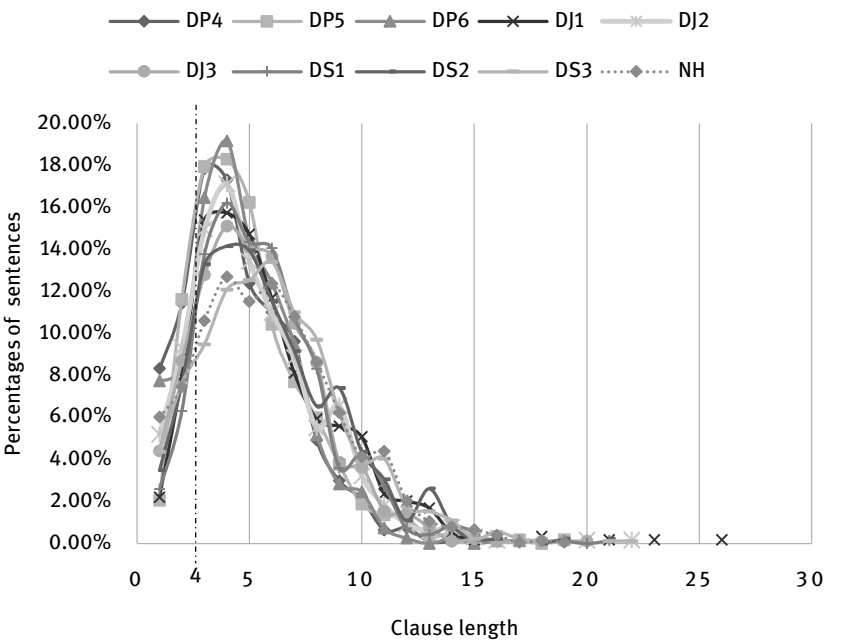


Fig. 6: Clause length distribution

From Fig. 7, we may find that the relationship between MDD and clause length presents a different trend from that of sentence length. For all grades, there is generally a linear correlation between clause length and MDD. Compared with the result of MDD and sentence length, the linear correlation between MDD and clause length is lowest in primary 4th grade, indicating no influence from clause length on MDD. However, the fitting result of the linear regression becomes consistently better starting from primary 6th grade. The high linear correlations in more grades suggest that compared with sentence length, MDD may perhaps be more heavily influenced by clause length variations. A linear growth by MDD

along with clause length early in primary 6th grade may indicate that DHH students, with their working memory gradually developing, may possibly start to produce more non-adjacent complex structures to satisfy more complicated communicative needs.

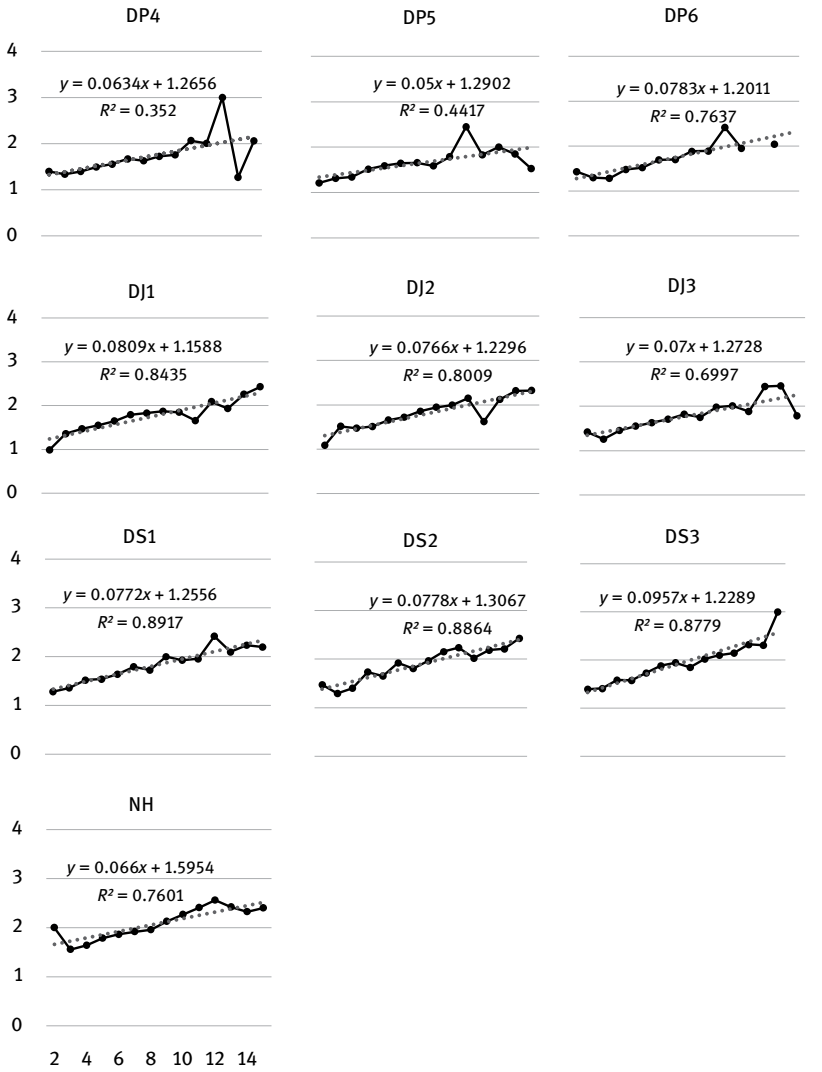


Fig. 7: MDD variation with different clause lengths in each grade from DHH and NH group. Numbers on the x-axis refer to the clause length, and those on the y-axis refer to the MDD.

4 General discussion

The present study examined the performance of dependency distance across different grades of DHH group by comparing with the performance of NH group from four perspectives, i.e., probability distribution, MDD variations, the influence by sentence length, and by clause length. We attempt to discover 1) the syntactic development in DHH students; 2) the way that DDM works on early syntactic performance. Four major findings are concluded.

Firstly, human language as a dynamic system can always reach a state of self-organization and self-regulation (Hřebíček & Altmann 1993), irrespective of the maturity of the language users. No matter how simple or fragmented the sentences are produced by language users, the inner integrity in language is ensured, so that even though misunderstandings and mistakes occur, the overall meaning can be understood by others. This study extends the applicability of Zipf-Alekseev function as a suitable model for distribution of length in dependency distance (Popescu et al. 2014). The Zipf-Alekseev function has been proposed and testified by Popescu et al. (2014) as a unified model of length of linguistic units at many levels, such as syllable length, word length and sentence length, and in many languages such as German, Chinese, Greek, etc. In this study, dependency distance as a length unit has also been corroborated to be distributed regularly in a long-tailed form. The regularity that the dependency distance distribution always abides by certain long-tailed distributions is partly attributed to the tendency of DDM (Liu et al. 2017).

Secondly, both the parameters of the DD distribution model and MDD can be effective indicators for linguistic complexity. Language changes over time, and it changes in the progress of use (Ellis 2008). Despite the fact that dependency distance distribution for DHH students generally conforms to the unified distribution laws, in this research, there is a general tendency for DHH students to approach the proficiency level of NH group as we found the smallest mean values of parameter α and longest MDDs in NH group as well as in advanced learning stages from DHH group. Besides, students from lower grades may consistently use short and simple syntactic relations, even in long clause lengths and sentence lengths. Although previous studies suggest that DHH students consistently lag behind hearing peers (see review: Luckner & Handley 2008), this study shows that at least from quantitative perspectives, the senior third grade students have achieved comparable performance with hearing students in syntactic complexity. With effective and long-term education, DHH students can improve their literacy level.

Thirdly, clause length may be a better predictor of MDD in developmental languages. In the present study, clause length was taken into consideration as a possible factor to predict MDD for the first time. The different trend for clause length from sentence length indicates a diverging pattern of influence on MDD. Although both sentence length and clause length are significantly related to MDD, clause length seems to be a better predictor of MDD in learners' languages. In addition, the finding that the majority of clauses produced by students are 4-word clauses seems to indicate a greater chance for clause length to be closely related with working memory capacity. Now that we propose DDM as the consequence of the limited storage costs, future studies may perhaps find more evidence of DDM by examining the relationship between clause length and MDD.

Lastly, this study extends the concept of DDM to developmental languages in a broader sense. We propose here that DDM may be the dominant rule in languages of the young children. In younger grades, especially in primary 4th and 5th grades, we found no influence of sentence length and clause length on mean dependency distance. As the language proficiency continues to develop along with cognitive maturity, more complex sentence structures need to be learned for diversified communicative needs. The production of such complex structures may need long dependency distance, naturally causing longer sentence length and clause length. Therefore, for those more skilled writers such as students from senior grades, they might use longer sentences with longer dependency distances. However, even though their MDDs have become more greatly influenced by sentence length or clause length, the MDDs have become constant when the sentence length reaches a certain value, which might suggest that dependency distance is constrained by working memory constraints (Liu et al. 2017). The pressure of DDM exists in the long-distance sentence structures as well. These findings yield potential implications as to what early syntactic development may be like. In the early learning stage, DDM seems to exert its strongest pressure and be manifested at the simplest form, represented by the dominant use of adjacent or short dependencies, to ensure the very basic communicative need. However, as communication cannot fulfill itself by merely using the simplest forms, one must continually seek a balance between the optimal communicative output and the minimal memory costs (Yan 2017). As cognitive capacity gradually becomes mature during adolescence (Luciana et al. 2005), the limited working memory can be liberated to produce longer dependencies while syntactic knowledge continues to develop with age (Slobin 1973).

5 Conclusion and implication

Using written language by deaf and hard-of-hearing students and by normal hearing students, the current study enriches the assumption of DDM in developing languages. It proposes that due to limited cognitive capacity and poor syntactic complexity by young learners, DDM has its strongest influence in the early languages as represented by the dominant use of adjacent dependency relations irrespective of sentence or clause length variations in the young learners' writings. Also, it further extends the use of quantitative analysis and dependency approach to the examination of syntactic development in deaf and hard-of-hearing students as special language users. Moreover, it proposes a potential influence of clause length on MDD. These findings may shed more light on the exploration of certain regularities in language acquisition and the uniqueness in language learners of a different type.

Although the possible influence of DDM in the early language output has been proposed in the present study, the relations between DDM, cognitive capacity and language proficiency is not clearly and directly examined. To get more evidence about the early manifestation of DDM and to find out the possible regularities of early syntactic development, it is also necessary to analyze learners' languages from other perspectives of dependency grammar approaches.

References

Altmann-Fitter. 2013. Altmann-fitter User Guide. The Third Version. Downloadable at: http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-UserGuide.pdf (2014-11-29).

Cowan, Nelson. 2005. *Working Memory Capacity*. New York: Psychology Press.

Cowan, Nelson. 2010. The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19* (1), 51–57.

De Villiers, Peter A. & Jill G. De Villiers. 1979. *Early Language*. Cambridge, MA: Harvard University Press.

Ellis, Nick C. 2008. The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal, 92* (2), 232–249.

Ferrer-i-Cancho, Ramon & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5* (2), 143–155.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (Eds.), *Image, Language, Brain* (pp. 95–126). Cambridge: The MIT Press.

Hřebíček, Luděk & Gabriel Altmann. 1993. Prospects of text linguistics. In Luděk Hrebicek & Gabriel Altmann (Eds.), *Quantitative Text Analysis* (pp.1–28). Trier: WVT.

Hudson, Richard. 1990. *English Word Grammar*. Oxford: Basil Blackwell.

Hudson, Richard. 1995. *Measuring Syntactic Difficulty*. Manuscript, London: University College.

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

Jiang, Jingyang & 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 Science, 50*, 93–104.

Jin, Huiyuan & Haitao Liu. 2016. Chinese writing of deaf or hard-of-hearing students and normal-hearing peers from complex network approach. *Frontiers in Psychology*, 7, 1777.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9* (2), 159–191.

Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.

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

Lu, Qian & Haitao Liu. 2016. Does dependency distance distribute regularly? *Journal of Zhejiang University (Humanities and Social Sciences),* 4, 63–76.

Luciana, Monica, Heather M. Conklin, Catalina J. Hooper & Rebecca S. Yarger. 2005. The development of nonverbal working memory and executive control processes in adolescents. *Child Development, 76* (3), 697–712.

Luckner, John L. & C. Michelle Handley. 2008. A summary of the reading comprehension research undertaken with students who are deaf or hard of hearing. *American Annals of the Deaf, 153* (1), 6–36.

Lyxell, Björn, Ulf Andersson, Erik Borg & Inga-Stina Ohlsson. 2003. Working-memory capacity and phonological processing in deafened adults and individuals with a severe hearing impairment. *International Journal of Audiology, 42*(sup1), 86–89.

Mitchell, Ross E. & Michael A. Karchmer. 2004. Chasing the mythical ten percent: Parental hearing status of deaf and hard of hearing students in the United States. *Sign Language Studies, 4* (2), 138–163.

Musselman, Carol & Gabriella Szanto. 1998. The written language of deaf adolescents: Patterns of performance. *The Journal of Deaf Studies and Deaf Education, 3* (3), 245–257.

Neophytou, Kyriaki, Marjolein van Egmond & Sergey Avrutin. 2017. Zipf's law in aphasia across languages: A comparison of English, Hungarian and Greek. *Journal of Quantitative Linguistics, 24* (2-3), 178–196.

Ninio, Anat. 2011. *Syntactic Development, Its Input and Output*. Oxford: Oxford University Press.

Ninio, Anat. 2014. Syntactic development: Dependency grammar perspective. In Patricia J. Brooks & Vera Kempe (Eds.), *Encyclopedia of Language Development*. London: SAGE Publications.

Nippold, Marilyn A., Tracy C. Mansfield, Jesse L. Billow & J. Bruce Tomblin. 2008. Expository discourse in adolescents with language impairments: Examining syntactic development. *American Journal of Speech-Language Pathology, 17* (4), 356–366.

Nippold, Marilyn A., Tracy C. Mansfield, Jesse L. Billow & J. Bruce Tomblin. 2009. Syntactic development in adolescents with a history of language impairments: A follow-up investigation. *American Journal of Speech-Language Pathology, 18* (3), 241–251.

Ouyang, Jinghui & Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*. DOI: 10.1080/09296174.2017.1373991

Popescu, Ioan-Iovitz, Karl-Heinz Best & Gabriel Altmann. 2014. *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Slobin, Dan I. 1973. Cognitive prerequisites for the development of grammar. In Charles A. Ferguson & Dan I. Slobin (Eds.), *Studies of Child Language Development* (pp.175–208), New York: Holt, Rinehart, & Winston.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.

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

Wolbers, Kimberly A., Hannah M. Dostal & Lisa M. Bowers. 2012. "I was born full deaf." Written language outcomes after 1 year of strategic and interactive writing instruction. *Journal of Deaf Studies and Deaf Education, 17* (1), 19–38.

Xue, Nianwen & Yaqin Yang. 2011. Chinese sentence segmentation as comma classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, 631–635. Association for Computational Linguistics.

Yan, Jingqi. 2017. Revisiting syntactic development in deaf and hearing children from a dependency approach: Comment on "Dependency distance: a new perspective on syntactic patterns in natural languages" by Haitao Liu et al. *Physics of Life Reviews, 21*, 207.

Hua Wang

Positional Aspects of Dependency Distance

Abstract: Based on Chinese dependency treebank PMT 1.0, the present study investigates the positional aspects of dependency distance (DD) quantitatively. Results show that (1) as the word position in the sentence increases, the tendency of mean dependency distance (MDD) in different sentence length groups shows striking similarity. The two longest MDDs generally are in the sentence-initial and sentence-final positions; (2) The consensus string (CS) and weighted consensus string (WCS) show some characteristics, which demonstrates again that human cognition plays an important role in affecting DD and dependency distance minimization is a universal tendency; (3) The distribution of DD in each sentential position can be captured by power law, which implies that something like a vertical structure of texts exists.

Keywords: dependency distance; position; quantitative linguistics

1 Introduction

Dependency grammar has been one of the hot topics in language study and many universal properties have been found based on the concept of dependency distance (hereafter referred to as DD), dependency relation and dependency direction (Liu et al. 2017). For instance, investigations on large-scale cross-linguistic materials indicate that DD minimization is a tendency in human language (Liu 2008; Futrell et al. 2015). DD and dependency direction are also verified to be a measure of vital importance in language typology (Liu 2008; Liu 2010; Liu et al. 2009). With the development of quantitative linguistics recently, many quantitative approaches have been employed to survey some properties in the framework of dependency grammar. Liu (2009) investigated the probability distribution of dependency relation based on a Chinese dependency treebank and results reveal that most of the investigated distributions can be fitted very well with a modified right-truncated Zipf-Alekseev distribution. Lu and Liu (2016) analyzed the probability distribution of DD in 30 languages and found that DD in human language may abide by a certain universal distribution pattern. Such publications abound. Nevertheless, to our best knowledge, few of

Hua Wang, Zhejiang University, Hangzhou, P.R.China, wanghuazju@163.com

https://doi.org/10.1515/9783110573565-011

them concern DD with its relation to position.

As a matter of fact, position of linguistic units is a fundamental and vital concept in quantitative linguistics. Investigations into position often show some clues for discovering universal properties of language. On one hand, for instance, many studies have revealed that there might be a relationship between the length of a linguistic unit and its sentential position. Uhlířová (1997) examined word position and word length in Czech and found that the first sentential position is the position of short words, whereas the final sentential position is the position of long words. Fan et al. (2010) quantitatively studied word length in individual sentential positions to test whether the tendency observed by Uhlířová (1997) holds in other languages. Results reveal that word length increases towards the end of the sentence across all the typologically different languages examined. Wang and Liu's (2014) investigation which explored the interrelationship between length, complexity of sentential constituents and their positions shows that within the sentence the longest and the most complex constituents tend to occur in the final position, and relatively shorter and less complex constituents tend to be in the initial position. Analogously, under the framework of dependency grammar each word (except for the root verb of the sentence) in different sentential positions is a dependent and has the distance to its governor, namely DD, which enables us to relate DD with its position and to treat DD as a special kind of length characteristics. Thus questions such as "is the DD in the sentence-initial position shorter than that in any other non-initial positions?" can be raised.

On the other hand, Zörnig et al. (2016) extended the concept of "consensus string" (hereafter referred to as CS) and "weighted consensus string" (hereafter referred to as WCS), terms that have been transferred from computational biology to linguistics recently (Zörnig & Altmann 2016), to investigate the positional occurrences of some linguistic units and frames. They found that "the text possesses also a 'vertical structure'" and "some law is working in the background" (Zörnig et al. 2016:134). The two concepts and the approach of investigating texts vertically can also be applied to study the positional aspects of DD and reveal some possible positional regularities. Definitions of CS and WCS are elaborated in section 2.2.

So our research question is: Are there some regularity between DDs and their sentential positions? To be specific, based on Chinese dependency treebank PMT 1.0 (Qiu et al. 2014) we intend to investigate the positional characteristics of DD quantitatively and answer the following questions:

a. How does the mean dependency distance (hereafter referred to as MDD) change as the word sentential position increases? Are there some

positions whose dependency relations tend to have longer or shorter DDs?

b. Do "consensus string" and "weighted consensus string" of DD show some characteristics?
c. How does DD distribute in each position in the sentence? What model or distribution can be used to describe the distribution of DD in each column? Do the parameters show some characteristics or regularities?

The current study is intended to show the "vertical structure" of DD and reveal some positional regularities of DD. Meanwhile, investigating DD vertically is a novel approach in quantitative linguistics. Thus the current investigation will enrich not only the study of dependency grammar but also the study of quantitative linguistics.

2 Materials and Methods

2.1 Dependency Grammar and Mean Dependency Distance in Word Positions

This section will introduce several basic concepts of dependency grammar.

In fact, there is no consensus on what dependency grammar is, but the following core properties of a syntactic dependency relation is generally accepted (Tesnière 1959; Hudson 1990; Liu 2009): it is a binary and asymmetrical relation with a direction between two linguistic units, in which one acting as the governor and the other as the dependent; it is labeled by a dependency type on top of the arc linking the two units. Fig. 1 presents the dependency analysis of a sample sentence in English.

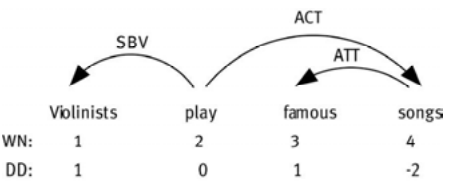


Fig. 1: Dependency structure of the sentence *Violinists play famous songs* (WN is word number and DD is dependency distance.)

As Fig. 1 shows, the arcs labeled with dependency types (see Appendix 1) direct from governors to dependents. For instance, in the pair *(play, songs)*, *play* is the governor, *songs* the dependent and their dependency relation is labeled as ACT. As for the DD, according to Liu et al. (2009), formally let $W_1...W_i...W_n$ be a word string in a sentence; for any dependency relation between the words W*m* and W*n*, if W*m* is a governor and W*n* is its dependent, then the DD between them is defined as the difference *m*–*n*; by this measure, the DD of adjacent words is 1. When *m* is greater than *n*, the DD is a positive number, which means the head follows the dependent; when *m* is smaller than *n*, the DD is a negative number and the governor precedes the dependent. For example, DD of the dependency relation (*play, songs*) in Fig. 1 is the difference 2–4, namely –2. However, in measuring DD the relevant measure is the absolute value of DD (ADD). That is to say, DD of the dependency relation (*play, songs*) is treated as 2 rather than –2 here.

In the present study, when talking about DD in different sentential positions, we refer to the distance of a word as a dependent to its governor in a given position. Thus it can be seen from Fig. 1 that DD in position 1 is 1 and in position 4 is 2. Position 2 is the position for root verb of the sentence with DD of 0.

MDD of a sentence has been used to discover linguistic regularities (Liu et al. 2009; Jiang & Liu 2015; Wang & Liu 2017). Liu et al. (2009) proposed a widely accepted method to measure the MDD of a sentence. The MDD of an entire sentence can be defined as:

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

Here *n* is the number of words in the sentence and DD_i is the DD of the *i*-th syntactic link of the sentence. Usually in a sentence there is one word (the root verb) without a head, whose DD is defined as zero. Thus MDD of the sentence in Fig. 1 is MDD = (1+1+2)/(4–1) = 1.33. This formula can also be used to calculate the MDD of a larger collection of sentences, such as a treebank:

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

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

Analogously, MDD in a sentential position *x* can be defined as

$$\text{MDD}(\text{ in a sentential position x in the sample}) = \frac{1}{n}\sum_{i=1}^{n}|DD_i| \tag{3}$$

Here *n* is the total number of words in a sentential position *x* in the sample, and DD_i is the DD of the *i*-th syntactic link in position *x* in the sample.

2.2 Definitions of Consensus String and Weighted Consensus String

In this section we will explain two concepts: "consensus string" (CS) and "weighted consensus string" (WCS).

According to Zörnig et al. (2016), a text can be considered vertically as a set of sequences in separate lines, i.e. a text can be transcribed into an array (1) as follows:

$$\begin{pmatrix} S_1{}^1 & S_2{}^1 & S_3{}^1 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ S_1{}^n & S_2{}^n & S_3{}^n & \dots \end{pmatrix} \tag{1}$$

where a sequence (a line or a string) represents a framing entity such as a sentence, a verse etc., and each S_j^i represents either symbols (such as word class) or numbers (such as word length) of the unit in the framing entity. That is to say, for example, if the framing entity is a sentence and S_j^i is the length of the unit of the sentence, namely word length, the array above might be concretized as array (2).

Since sequences like sentences are not necessarily of equal length, one can add zeros to bring them to the same lengths as we did in array (2). In an array like (1) and (2), Zörnig et al. (2016) define a CS as a sequence $t = (t_1,\dots,t_n)$ which is as close as possible to the strings, and one possibility to define $t = (t_1,\dots,t_n)$ is to set t_j equal to one of the most frequent elements of the j-th column. Taking array (2) as an example, the most frequent element of the 1st column is 1, and the 2nd column 2, the 4th column 4, the 6th column 2 and the 8th column 0. In the 3rd and 7th column each element can be seen as the most frequent one since all of them occur once and in the 5th column there exist two most frequent elements, 3 and 5, so each of them could be the CS of this column.

Column	1	2	3	4	5	6	7	8
S^1	1	2	4	3	5	3	2	0
S^2	1	2	2	4	3	2	1	3
S^3	2	1	3	4	3	1	0	0
S^4	1	2	5	4	5	2	3	0

(2)

From above we can see that CS of the strings are not unique. Therefore, one possible CS of (2) might be:

CS = (1, 2, 4, 4, 3, 2, 1, 0),

and another possible CS might be:

CS = (1, 2, 2, 4, 5, 2, 2, 0).

The definition of CS implies that CS actually minimizes the average distance to the given strings (Zörnig & Altmann 2016).

As for the WCS, "Let $F(j)$ be the largest frequency of an element in column j and let $N(j)$ denote the number of elements in column j"(Zörnig et al. 2016:2), then the WCS is defined as a sequence:

$$\left(\frac{F(j)}{N(j)}\right)_{j=1, 2, 3.....}$$

Different from CS, WCS is always a uniquely determined numeric sequence. In array (2), the most frequent element in column 1 is 1 with the frequency of 3 and the sum of the elements is 4, thus $F(1) = 3$ and $N(1) = 4$. The WCS value in column 1 is 3/4. Column 8 contains only 1 element (others are 0 which are added to get the same length) and the most frequent element is 3 with the frequency of 1, so $F(1) = 1$ and $N(1) = 1$ and the WCS value in column 8 is 1/1, and so on. The WCS of array (2) is therefore

WCS = (3/4, 3/4, 1/4, 3/4, 2/4, 2/4, 1/3, 1/1).

But all complete strings like array (1) and array (2) are not needed in practice. A frequency table of the string matrix is often sufficient to calculate CS and WCS. The frequency table of array (2) above is given in Tab. 1, where C is the column or word position and L is the length of word.

Tab. 1: The frequency table of array (2)

L \ C	1	2	3	4	5	6	7	8
1	3	1	0	0	0	1	1	0
2	1	3	1	0	0	2	1	0
3	0	0	1	1	2	1	1	1
4	0	0	1	3	0	0	0	0
5	0	0	1	0	2	0	0	0
sum	4	4	4	4	4	4	3	1

A string matrix above facilitates not only the investigation of CS and WCS of the strings but also the rank-frequency distributions of the individual columns.

2.3 The Treebank and Data Processing Tools

The Chinese dependency treebank we used here is part of the Peking University Multi-view Chinese Treebank version 1.0 (PMT 1.0), which is freely available for any researcher. It contains 14,463 sentences and 336,000 words. This main corpus is based on the Peking University People's Daily Corpus and the texts are from all the articles of People's Daily from January 1st to January 10th, 1998 (Qiu et al. 2014; Qiu et al. 2015).

Tab. 2 exhibits some information of a sentence example extracted from PMT 1.0, among which sn (sentence number), wn (word number), sl (sentence length), wt (word type), dep (dependency type) and gn (governor number) have already been provided and dd (dependency distance) and add (absolute dependency distance) are added into the treebank through operation (dd = gn−wn and add =|gn−wn|).

In the phase of data preprocessing, the root verbs of the sentences with DD of 0 and punctuations which are not considered as words are deleted. Then with the help of software *R* we obtained a frequency table of the string matrix of word position and DD. Tab. 3 shows part of the matrix. The fittings were done by the fitting software Nonlinear Regression and Curve fitting (Liu 2017).

Tab. 2: A sentence example in PMT 1.0

sn	wn	sl	wn	wt	dep	gn	dd	add
wn	1	15	在	p	ADV	11	10	10
sl	2	15	这	r	ATT	4	2	2
wn	3	15	一	m	NUM	4	1	1
wt	4	15	年	q	ATT	5	1	1
dep	5	15	中	f	POB	1	−4	4
gn	6	15	,	w	PUN	11	5	5
dd	7	15	中国	ns	DE	8	1	1
add	8	15	的	u	ATT	10	2	2
19	9	15	外交	n	ATT	10	1	1
1	10	15	工作	n	SBV	11	1	1
15	11	15	取得	v	HED	0	−11	11
在	12	15	了	u	MT	11	−1	1
p	13	15	重要	a	ATT	14	1	1
ADV	14	15	成果	n	VOB	11	−3	3
11	15	15	。	w	PUN	11	−4	4

(***Note.*** The translation of the Chinese sentence is:
在 这一 年中，中国的 外交 工作 取得了 重要 成果 。
This year, Chinese diplomatic work has achieved important outcomes.)

Tab. 3: Part of the frequency table of the string matrix of word position and DD (WP is word position in the sentence and DD is the dependency distance.)

DD \ WP	1	2	3	4	5	6
1	5762	6033	6109	6289	5667	5346
2	1702	1480	2017	1980	1975	1848
3	911	996	887	1110	1068	986
4	787	639	524	477	812	663
5	491	362	346	318	325	585
6	393	270	214	240	249	225

3 Results and Discussions

3.1 Variation Tendency of the Mean Dependency Distance with the Increase of Word Position

DD is affected by various factors and one of the often mentioned is sentence length (Jiang & Liu 2015). Mixing the sentences with different lengths together may raise risks and result in distorted results (Ferrer-i-Cancho & Liu 2014); therefore, it is desirable to take sentence length into account and group sentences with the same length together to examine the variation tendency of MDD in different sentential positions. Results are shown visually in Fig. 2. It's noteworthy that the sentence lengths in PMT 1.0 range from 1 to 198, but sentences with length shorter than 4 or longer than 35 are excluded, since the former have none or only one positional variation and the latter are relatively small in quantity, which might skew the result if included. Meanwhile, for the reason that the sentence-final punctuations have been deleted, each sentence of length n has n–1 word positions.

Fig. 2 shows that as the sentential position increases, the tendency of MDD in different sentence length groups shows striking similarity. Generally, the sentence-initial and sentence-final word positions share the two longest MDDs among all the positions, and MDDs in all the non-initial and non-final positions are much shorter. It's interesting that MDD starts to decrease slowly in position 2 and reaches the smallest value in the second last position. These results indicate that some word positions in the sentence, especially the sentence-initail and sentence-final positions, do show peculiar charactertistics in MDD.

One may wonder what factors are resposible for this. Our first hypothesis is that it might be attributed to the distribution of dependency relations (or dependency type) in different positions. Previous studies in Chinese dependency treebank indicate that some dependency types tend to be much longer than others and several possible causes that make the differences are also provided (Liu 2007a).

For testing our speculation, we extracted the ten most frequent dependeny types and calculated their MDDs in sentence-initial and sentence-final positions. Fig. 3 and Fig. 4 show the distribution of the ten most frequent dependency types in sentence-initial and sentence-final positions respectively, and Tab. 4 lists MDDs of different dependency types, in which action object is referred to as ACT, appositive element APP, complement CMP, imaginative IMA, other coordination element COO, modifier of 的 (of) DE, independent clause IC,

left additive LAD, modality and time MT, number NUM, prepositional object POB, postpositional quantity QUC and right additive RAD (Appendix 1 lists all the 32 dependency types used in PMT 1.0).

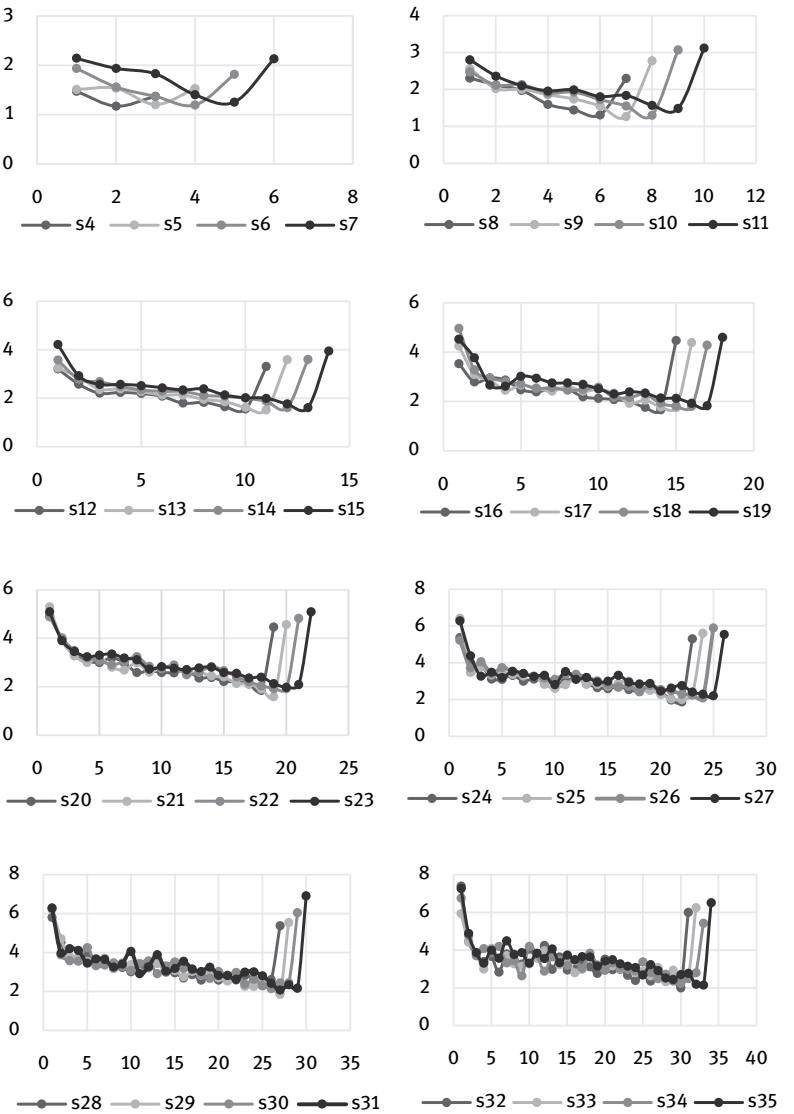


Fig. 2: Variation tendency of MDD with the increase of sentential positions in different sentence length groups

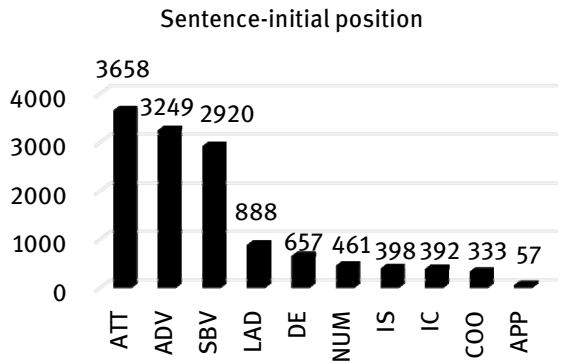


Fig. 3: Distribution of the ten most frequent dependeny types in sentence-initial position

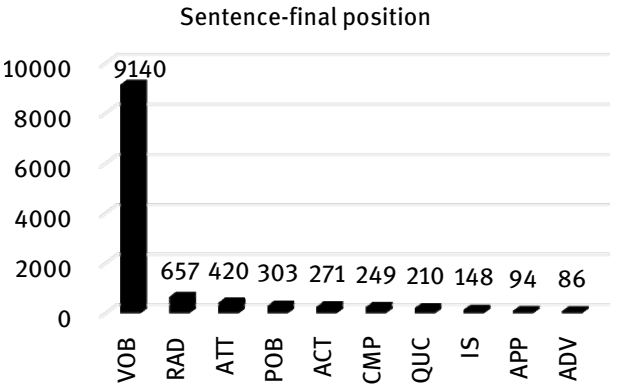


Fig. 4: Distribution of the ten most frequent dependeny types in sentence-final position

In sentence-initial position, though ATT with MDD of 1.58 has the largest proportion, its frequency is close to ADV and SBV, whose MDDs are 4.23 and 2.99 respectively. It's noteworthy that IS and IC with strikingly high MDDs of 10.38 and 9.84 occur 790 times in total, which might be another reason for the longer MDD in sentence-initial position. The situation in sentence-final position is less complicated since two-thirds of the dependency relations are VOB with a longer MDD of 4.67, so it's not surprising that sentence-final position has such a long MDD. Those results suggest that the longer MDD in sentence-initial and sentence-final positions can be partly ascribed to the distirbution of dependency relations, but other reasons still need to be explored.

Tab. 4: MDDs of different dependency types

Dependency type	MDD	Dependency type	MDD
IC	10.38	QUC	2.43
IS	9.84	QUN	2.37
ACT	5.61	DE	1.74
COO	4.77	APP	1.66
VOB	4.67	ATT	1.58
ADV	4.23	RAD	1.43
POB	3.75	NUM	1.32
LAD	3.36	CMP	1.22
SBV	2.99	MT	1.08

If we merely put emphasis on the MDD in sentence-initial and sentence-final positions, we can find that as the sentence length increases, the MDDs in the two positions increase as well as shown in Fig. 5. Furthermore, the relationship between sentence length and MDDs in the two positions are most likely to be linear, ie. $y= a+bx$, where x stands for sentence length and y the corresponding MDD. The linear relationships are presented in Fig. 6.

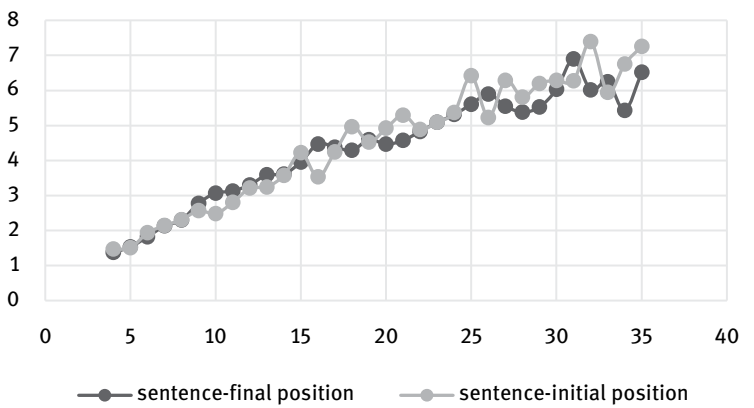


Fig. 5: MDDs in sentence-initial and sentence-final positions with the increase of sentence length

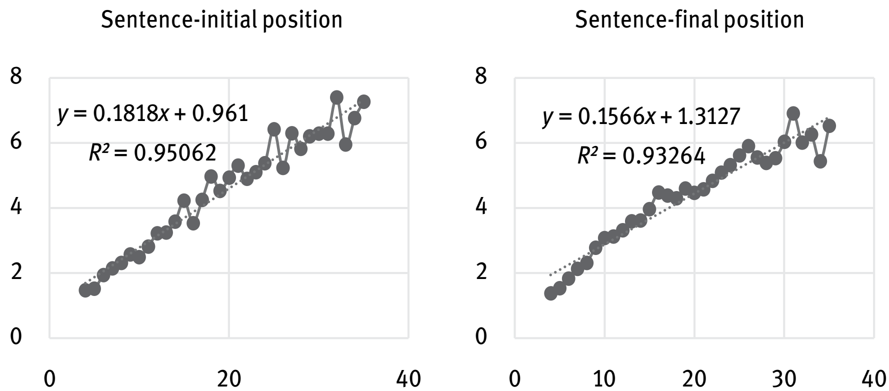


Fig. 6: The linear relationships between MDD and the two positions respectively

According to Fig. 6, the slopes *b* in initial and final positions are very small, which suggests that MDDs in the two positions only change within a very small range. Meanwhile, the two slopes are similar (0.182 versus 0.157), which indicates that though MDDs in the two positions increases with sentence length, their acceleration speeds are both slow (Jiang & Liu 2015).

3.2 CS and WCS of Dependency Distance

According to the string matrix of word position and DD (due to the limited space, the whole matrix cannot be presented and part of it is shown in Tab. 3), we obtained the unique CS,

CS = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1……),

which means that in each position or in each column the most frequent DD is 1. According to the definition, CS is not unique but with many possible forms, since there might be more than one most frequent elements in a position or a column. However, the CS here is unique for the reason that the element with DD of 1 is absolutely the most frequent one in every position.

CS is defined as a string that is as close as possible to the given strings and minimizes the average distance to them (Zörnig & Altmann 2016). That is to say, a sequence or a sentence with DD of 1 in every word position is as close as possible to all the sentences, though it's only an ideal situation. Such a CS suggests again that DD minimization is a universal tendency.

The WCS of DD in word position is

WCS =(0.4382 0.5296 0.5329 0.5413 0.4994 0.4992 0.5060 0.5056 0.5010
0.4911 0.4996 0.4875 0.4885 0.4889 0.4890 0.4940 0.4803 0.4892
0.4882 0.4845 0.4895 0.4874 0.4792 0.4858 0.4816 0.4763 0.4907
0.4880 0.4676 0.4859 0.5015 0.4902 0.4544 0.4827 0.4977 0.4867
0.4800 0.4671 0.4876 0.4943 0.4993 0.4668 0.5246 0.4491 0.4691
0.4709 0.4772 0.4743 0.5059 0.4822 0.4562 0.5302 0.4583 0.4947
0.5153 0.4700 0.4943 0.4910 0.4674 0.4800 0.4980 0.4690 0.4339
0.5323 0.4520 0.4487 0.5067 0.5714 0.4776 0.4874 0.4815 ...),

and the WCS is also presented visually in Fig. 7. The positions higher than 71 were not included due to the small quantity of sentences. Since the most frequent element in every position is the dependency relation with DD of 1 which is also called adjacent dependency, we can interpret WCS as how the probability of adjacent dependencies changes with the increase of word position in sentence.

Adjacent dependency is of vital importance in the distribution of DD. Based on twenty languages, Liu (2008) found that about half of the dependency relationships in human languages are adjacent dependencies. Based on five Chinese treebanks with different annotating schemes, genres and average sentence lengths, Liu et al. (2009) further found that the percentages of adjacent dependencies of Chinese range from 47.9% to 56.6% with an average of 53.06%. The percentages of adjacent dependencies in Chinese in Jiang and Liu's (2015) study are kept stable within 55% to 64% in different sentence lengths and the average percentage is 59.6%. Wang and Liu's (2017) study shows that the percentage of adjacent dependencies in English tends to decrease with an increase in sentence length. These above investigations show that the distribution of adjacent dependencies has its regularities though it may differ a little by the influence of language, sentence length or other factors. Then our question is whether regularities exist in the distribution of adjacent dependencies as the word position increases. We have mentioned before that the WCS value in the present study actually is the probability of adjacent dependencies.

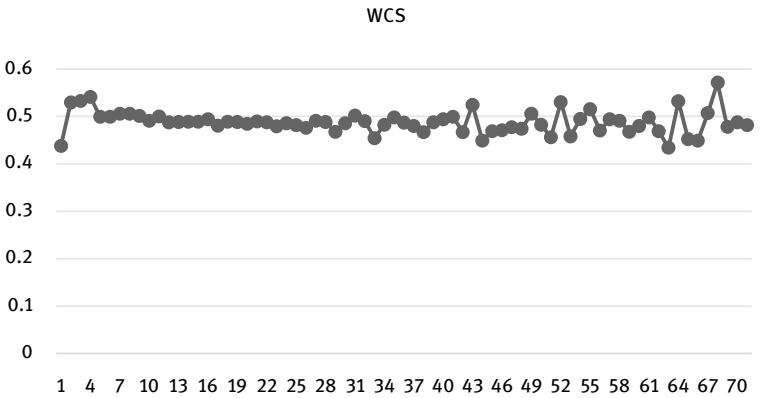


Fig. 7: The WCS of DD in word position

Therefore, from Fig. 7 we can see that the lowest probability 0.4382 appears in position 1, which also provides another explanation for the longest MDD in position 1, because the number of adjacent dependencies influences the MDD (Liu 2008). In position 2, 3 and 4 the probabilities increase sharply with the value of 0.5296, 0.5329 and 0.5413 respectively. Starting from position 5 to position 30, the probabilities are kept stable around 0.50. In the position higher than 30, the probabilities have a violent fluctuation, which might be ascribed again to the small numbers of sentences. Nevertheless, no matter how the probabilities fluctuate, they are never out of the range between 0.4 to 0.6. Further, the probability of fluctuating around 0.5 is consistent with Liu (2008) and Liu et al.'s (2009) results aforementioned and is also in line with Wang and Liu's (2017) results that adjacent dependencies always account for approximately half of the values in DD distribution of different genres. In short, the present study suggests that a regularity may exist in the distribution of adjacent dependencies in word position.

The unique CS = (1,1,1,……) and the value of WCS demonstrate once more that human cognition plays an important role in affecting DD and DD minimization is a universal tendency. Human's memory capacity is limited, and the shorter the DD, the lower the cognitive load; therefore, a high percentage of adjacent dependencies is preferred to minimize the cognitive cost of human brain (Liu 2008; Ferrer-i-Cancho 2013).

3.3 The Distribution of Dependency Distance in Each Sentential Position

This section addresses the distribution of DD in each sentential position. When investigating positional occurrence in texts, Zörnig et al. (2016) conjectured a hypothesis that "each column in the string matrix has its specific frequency distribution" and "the confirmation of this conjecture implies that something like a vertical structure of texts exists" (Zörnig et al. 2016:4). As a result, for some measured values (for instance, word length) Zörnig et al. (2016) found the Hyperpoisson distribution and some modified cases, and for rank-frequencies they found the well-known zeta distribution, which corroborates the proposed hypothesis.

For the present study, we can conjecture that each column or position in the string matrix of DD as shown in Tab. 3 has its specific frequency distribution, but what kind of distribution or model can be used?

In the previous studies of the DD distribution, various models and distributions have been employed. Liu (2007b) investigated the probability distributions of DDs in Chinese and results reveal that the data can be well captured by the right truncated zeta distribution. Lu and Liu's (2016) study reveals that DD distribution patterns all fit a mixed exponential and power law distribution. Wang and Liu (2017) used right truncated zeta distribution, right truncated waring distribution and exponential distribution to capture the distributions of DDs of all sentence lengths and genres. Jiang and Liu (2015) fitted the power law distribution to the relationship between sentence length and MDD. Ferrer-i-Cancho and Arias' (2013) investigation on a Catalan treebank also suggests an almost power-law dependency between sentence length and MDD. Though presented in different forms, in essence the distribution models are no other than Zipf-like laws (power law distribution) or exponential distribution.

Therefore, we assume that the distribution of DD in each sentential position can be captured by the power law distribution:

$$y = ax^{-b}, \tag{4}$$

where x represents DD and y represents its frequency. The fitting results are presented in Tab. 5 and Fig. 8. Due to the limitation of space, only the data of the first six positions are displayed here (see complete fitting results in Appendix 2).

The model fittings to power law distribution are excellent with all the R^2 values higher than 0.90. On one hand, it corroborates our assumption that DD in each position has its specific frequency distribution, which verifies Zörnig et

al. 's (2016) statement that something like a vertical structure of texts does exist from a new perspective. On the other hand, it demonstrates that we can find Zipf-like laws no matter in the overall distribution of DD, the distributions of DD in different genres, the distributions of DD in sentences with different lengths or the distributions of DD in different positions. The finding of this kind of long-tailed distribution in DD shows once again the tendency of DD minimization, since in long-tailed distributions large quantities of DD are shorter ones while longer ones are rare.

Tab. 5: Fitting results of power law distribution to DD's distribution in the first six positions

position	*a*	*b*	R^2
1	5708.2866	1.5702	0.9962
2	5997.1444	1.7864	0.9969
3	6133.6048	1.7135	0.9990
4	6308.9449	1.7179	0.9987
5	5696.9045	1.5867	0.9973
6	5369.0103	1.5815	0.9974

Now we will discuss parameter *a* and parameter *b*. Fig. 9 shows the relationship between position and values of *a* and *b* respectively. The value of parameter *a* gradually increases from position 1 to position 4 and then decreases as the position increases. Parameter *a* actually is the estimated frequency of DD with 1 in each position, which means that the variation of parameter *a* is the frequency variation of DD with 1. When it comes to parameter *b*, we found that almost all of them change in the range from 1.5 to 1.8. To be specific, except for position 2, 3 and 4 that have larger values, *b* is very stable from position 1 to position 40 with the value around 1.5. The violent fluctuation of parameter *b* after position 40 might be ascribed to the relatively small numbers of sentences again.

Compared with the value of WCS presented in Fig. 7, an interesting phenomenon is found and Fig. 10 graphically shows that the two lines are almost parallel though the variation of WCS is sharper than that of parameter *b*. A Pearson correlation test was performed to explore their relationship, with Pearson correlation = 0.850, $p < 0.001$, showing that WCS (or the percentage of adjacent dependencies in a position) and parameter *b* are positively correlated. Further investigation is needed to explain why such correlation exists.

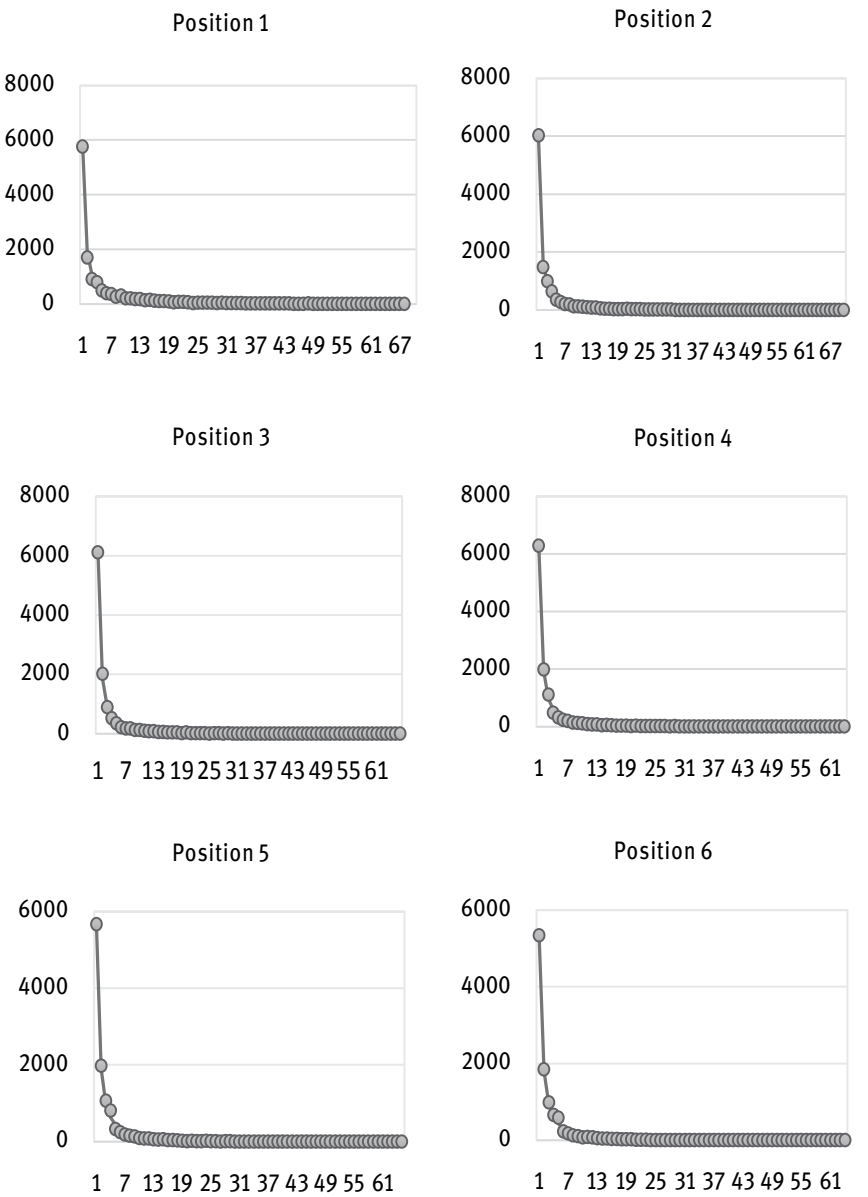


Fig. 8: Fitting the power law to the distribution of DD in the first six positions (The solid line is the model fit and the small circles are the observed values.)

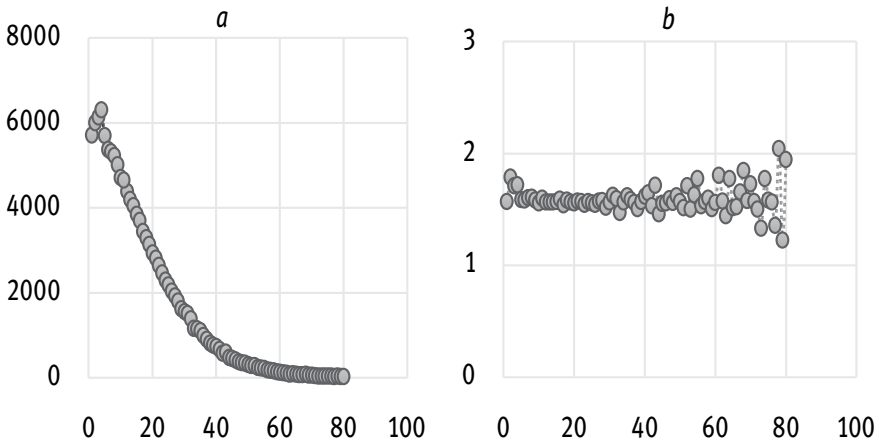


Fig. 9: Parameters *a* and *b* in the fitting of power law (x-axis represents the position and y-axis represents the values of *a* and *b*.)

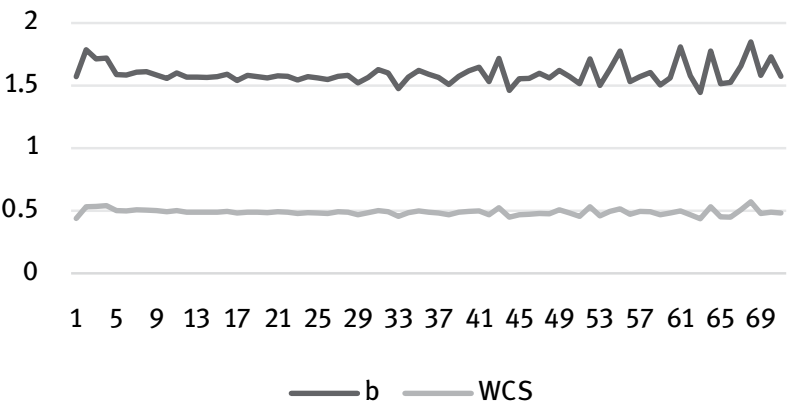


Fig. 10: The relationship between WCS and parameter *b* (The upper line represents parameter *b* and the lower line represents WCS.)

4 Conclusions

There are some regularities in the positional aspects of DD.

Firstly, as the word position in sentence increases, the tendency of MDD in different sentence length groups shows striking similarity. The two longest MDDs are generally in the sentence-initial and sentence-final positions. Then from position 2 MDD starts to decrease slowly and reaches the smallest value in the second last position. The distribution of dependency relaitons in different positions is found to be partly responsible for the longer MDD in sentence-initial and sentence-final positions.

Secondly, the CS and WCS of DD exhibit some characteristics. The CS of DD is unique with CS = (1, 1, 1 ...). The value of WCS in the present study is actually the probability of adjacent dependencies. Position 1 has the lowest probability of 0.4382, which also explains the longest MDD in position 1. Then the probabilities increase sharply in position 2, 3 and 4; starting from position 5 to position 30 the probabilities are kept stable around 0.50. Nevertheless, no matter how the probabilities fluctuate, it is always confined to 0.4~0.6. This study suggests that a regularity may exist in the distribution of adjacent dependencies in different word positions. These findings demonstrate once again that human cognition plays an important role in affecting DD and DD minimization is a universal tendency.

Thirdly, the distribution of DD in each sentential position can be captured by power law distribution, which verifies Zörnig et al.'s (2016) statement that something like a vertical structure of texts does exist from a new perspective. Parameter *b* in the power law shows some regularities. On one hand, almost all of them change in the range from 1.5 to 1.8. On the other hand, it's interesting that WCS and parameter *b* are positively correlated.

However, the present study only focuses on Chinese. Can the results we obtained from Chinese be generalized to other languages? In addition, some results are not fully explained. For instance, why are the WCS and parameter *b* in power law positively correlated? So further investigations are still needed.

References

Fan, Fengxiang, Peter Grzybek & Gabriel Altmann. 2010. Dynamics of word length in sentence. *Glottometrics, 20*, 70–109.

Ferrer-i-Cancho, Ramon. 2013. Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics, 25*, 1–21.

Ferrer-i-Cancho, Ramon & Marta Arias. 2013. Non-linear regression on dependency trees. *Lecture on Complex and Social Networks* (2013–2014). http:// www.lsi.upc.edu/wCSN/lab/session3.pdf (accessed 29 November 2014).
Ferrer-i-Cancho, Ramon & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5* (2), 143–155.
Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS, 112*(33), 10336–10341.
Hudson, Richard. 1990. *English Word Grammar*. Oxford: Basil Blackwell.
Jiang, Jingyang & 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 Science, 50*, 93–104.
Liu, Haitao. 2007a. Dependency relations and dependency distance: A statistical view based on treebank. In *Proceedings of the 3rd International Conference on Meaning–Text Theory*, 269–278. Klagenfurt, Austria.
Liu, Haitao. 2007b. Probability distribution of dependency distance. *Glottometrics, 15*, 1–12.
Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.
Liu, Haitao. 2009. Probability distribution of dependencies based on a Chinese dependency treebank. *Journal of Quantitative Linguistics, 16*(3), 256–273.
Liu, Haitao, Yiyi Zhao & Wenwen Li. 2009. Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics, 45(*4), 495–509.
Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120*(6), 1567–1578.
Liu, Haitao. 2017. *An Introduction to Quantitative Linguistics*. Beijing: The Commercial Press.
Lu, Qian & Haitao Liu. 2016. Does dependency distance distribute regularly*? Journal of Zhejiang University (Humanities and Social Sciences), 46*(4), 63–76.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.
Qiu, Likun,Yue Zhang, Peng Jin & Houfeng Wang. 2014. Multi-view Chinese treebanking. In *Proceedings of the 25th International Conference on Computational Linguistics*, 257–268. Dublin, Ireland.
Qiu, Likun, Peng Jin & Houfeng Wang. 2015. A multi-view Chinese treebank based on dependency grammar. *Journal of Chinese Information Processing, 29*(3), 9–15.
Uhlírová, Ludmila. 1997. Length vs. order: Word length and clause length from the perspective of word order. *Journal of Quantitative Linguistics, 4*(1–3), 266–275.
Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.
Wang, Hua & Haitao Liu. 2014. The effects of length and complexity on constituent ordering in written English. *Poznań Studies in Contemporary Linguistics, 50*(4), 477–494.
Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences, 59*, 135–147.
Zörnig, Peter & Gabriel Altmann. 2016. *Consensus Strings*. Lüdenscheid: VAM-Verlag.
Zörnig, Peter, Kamil Stachowski, Ioan-Iovitz Popescu, Tayebeh Mosavi Miangah, Ruina Chen & Gabriel Altmann. 2016. *Positional Occurrences in Texts: Weighted Consensus Strings*. Lüdenscheid: RAM-Verlag.

Appendix

Appendix 1: The 32 dependency categories used PMT 1.0

Tag	Dependency type
ACT	action object
ADV	adverbial
APP	appositive element
ATT	attribute
CMP	complement
COO	other coordination element
COS	share-right-coordination element
DE	de (modifier of 的(special function word))
DEI	dei (modifier of 得(special function word))
DI	di (modifier of 地(special function of word))
FOC	focus
HED	root of a sentence
IC	independent clause
IOB	indirect object
IS	independent structure
ISC	non-shared independent structure
LAD	left additive
MT	modality and time
NUM	number
POB	propositional object
PUN	punctuation
PUS	cross-clause punctuation
QUC	post-positional quantity
QUCC	Non-shared post positional quantity
QUN	quantity
RAD	right additive
RADC	non-shared right additive
RED	reduplicate element
SBV	subject
TPC	topic
VOB	direct object
VV	serial verb construction

Appendix 2: Fitting results of power law distribution to DD's distribution in each position

position	*a*	*b*	R^2
1	5708.2866	1.5702	0.9962
2	5997.1444	1.7864	0.9969
3	6133.6048	1.7135	0.9990
4	6308.9449	1.7179	0.9987
5	5696.9045	1.5867	0.9973
6	5369.0103	1.5815	0.9974
7	5315.5421	1.6052	0.9979
8	5222.2235	1.6097	0.9987
9	5010.4261	1.5818	0.9983
10	4714.7574	1.5568	0.9982
11	4650.2838	1.6004	0.9986
12	4384.6581	1.5675	0.9990
13	4196.8838	1.5677	0.9988
14	4047.8200	1.5645	0.9989
15	3849.6648	1.5700	0.9989
16	3701.7265	1.5899	0.9992
17	3437.4916	1.5401	0.9986
18	3296.5616	1.5814	0.9991
19	3136.2564	1.5683	0.9987
20	2938.7971	1.5593	0.9990
21	2809.7089	1.5763	0.9991
22	2644.6650	1.5726	0.9987
23	2469.2994	1.5431	0.9988
24	2296.8405	1.5686	0.9989
25	2180.2156	1.5585	0.9990
26	2036.8810	1.5454	0.9989
27	1927.8768	1.5738	0.9992
28	1799.2318	1.5804	0.9983
29	1621.5333	1.5198	0.9987
30	1557.5096	1.5645	0.9989
31	1508.2688	1.6266	0.9995
32	1382.0833	1.5980	0.9990
33	1160.5500	1.4735	0.9976
34	1144.9590	1.5673	0.9982

position	*a*	*b*	R^2
35	1100.6156	1.6207	0.9992
36	991.8842	1.5889	0.9991
37	902.3069	1.5639	0.9978
38	816.2836	1.5084	0.9972
39	769.3997	1.5718	0.9987
40	738.9746	1.6177	0.9989
41	669.9418	1.6467	0.9944
42	579.2897	1.5303	0.9988
43	598.3086	1.7144	0.9989
44	477.2185	1.4605	0.9953
45	456.5914	1.5547	0.9978
46	421.6598	1.5581	0.9986
47	374.7849	1.5962	0.9964
48	350.5395	1.5611	0.9986
49	346.0522	1.6205	0.9947
50	311.8533	1.5695	0.9973
51	282.0808	1.5134	0.9977
52	289.3901	1.7117	0.9978
53	236.6033	1.5012	0.9958
54	232.1028	1.6309	0.9976
55	217.4108	1.7758	0.9949
56	180.7664	1.5317	0.9973
57	173.5717	1.5714	0.9965
58	163.7814	1.6026	0.9950
59	138.0856	1.5050	0.9891
60	133.0637	1.5609	0.9956
61	122.0495	1.8073	0.9839
62	105.2706	1.5765	0.9827
63	82.2061	1.4438	0.9852
64	98.5284	1.7761	0.9934
65	78.6463	1.5134	0.9748
66	69.8915	1.5231	0.9925
67	75.9811	1.6557	0.9938
68	80.1975	1.8482	0.9956
69	64.2650	1.5787	0.9930
70	57.5209	1.7297	0.9870

position	*a*	*b*	R^2
71	51.7466	1.5720	0.9910

Jinlu Liu*, Gaiying Chai

Dependency Distance and Direction of English Relative Clauses

Abstract: Based on dependency syntactic treebanks, the present study focuses on relative clauses (RCs) and explores their relations between dependency distance (DD), dependency direction, the embedding position, and the length of RC. It was found that: (1) the probability distributions of the DD of RCs aren't influenced by their embedding positions; (2) the DD of RC embedded in the object position is relevant to the length of the RC; (3) the longer the RC is (≥10), the more likely the dependency relations within the RC present a head-initial tendency; (4) the embedding positions of RCs in main clauses don't influence their dependency directions and dependency directions of RCESs and RCEOs are almost head-initial; (5) the changes of dependency relations of main clauses are related to the embedding position of relative clauses; (6) the embedded RCs increase the MDDs of their main clauses significantly.

Keywords: dependency distance; dependency direction; dependency treebank; relative clause; English complex sentence

1 Introduction

According to Quirk et al. (1985: 1047), English sub-clauses include relative clauses, nominal clauses and adverbial clauses. It is because of their syntactic complexity (Gibson 1998, 2000; Hawkins 1994, 2004) and typological richness and value (Keenan & Comrie 1977) that the processing or comprehension difficulty of relative clause has been an important research topic in the last few decades.

The systematic study of syntactic structure of relative clauses can be traced back to the following three hypotheses. Perceptual Difficulty Hypothesis (PDH) (Kuno 1974) is concerned with the position of relative clauses and explores the implicational relationship between the two types of relative clauses, which include relative clauses in subject and object position of main clauses, holding

Jinlu Liu, Zhejiang University, Hangzhou, P.R.China, hardeningwings@163.com
Gaiying Chai, Zhejiang Gongshang University, Hangzhou, P.R.China

https://doi.org/10.1515/9783110573565-012

that due to the limitations of short-term memory, the relative clause embedded in subject position of main clause is more difficult to process than the one embedded in object position. For example, according to Kuno (1974: 119), "*The cheese that the rat that the cat chased ate was rotten*." is more difficult to process than "*The cat chased the rat that ate the cheese that was rotten*." The hypothesis has been confirmed and supported by many studies (Romaine 1984; Hamilton 1994; Bates et al. 1999; Izumi 2003; Xiao & Lv 2005). Noun Phrase Accessibility Hierarchy (NPAH) (Keenan & Comrie 1977), based on the typological markedness obtained from the comparative researches in many different types of languages, is concerned with the type of relative clauses. NPAH holds that the subclauses pose different levels of processing difficulty according to the accessibility hierarchy: SU (subject) > DO (direct object) > IO (indirect object) > OP (object of preposition) > GEN (genitive) > OCOMP (object of comparison) (">" indicates "to be more accessible than ..."). The hypothesis has been confirmed and supported partly by many scholars (Diessel & Tomasello 2005; Marefat & Rahmany 2009; Hou 2011). Subject Object Hierarchy Hypothesis (SOHH) (Hamilton 1994) is concerned with the cognitive difficulty of relative clauses, which integrates the above two hypotheses and predicts the difficulty levels of cognitive processing of relative clauses according to the amount of processing discontinuity, which is partly supported by some scholars (Warren & Gibson 2002; Izumi 2003; Reali & Christiansen 2007; Street 2017).

In fact, PDH and SOHH only pay close attention to the position and type of relative clauses, but it has not been explored how a relative clause has an effect on its main clause in the process of forming a complex sentence. Since the comprehension difficulty of main clauses, in which different relative clauses are embedded, might be also different, it is of much significance to compare relative clauses in terms of the comprehension difficulty of the corresponding complex sentences.

In addition, the research is rare on the relation between the length of a relative clause and its comprehension difficulty. Liu (2008: 178) regarded dependency distance as a metric reflecting the comprehending difficulty, thus we could investigate the relationship between the length of relative clause and dependency distance. By comparing the difference of dependency distance of relative clauses with different length, we can explore the effects of the length of relative clause on the comprehension difficulty of English relative clause.

Dependency distance may reflect the direction of a dependency, which can serve as an indicator for word-order typology (Liu 2010: 1567). We can resort to dependency direction to make clear whether different positions of relative clauses have different distributions of dependency relations. Besides, depend-

ency direction can be also applied to investigate the effects of relative clause on dependency relation of main clause. The answers to the above questions may be conducive to our comprehending the difference of syntactic feature of English relative clauses.

From the perspectives of dependency grammar and quantitative linguistics, our present study attempts to answer the following questions:

1) Is the probability distribution of dependency distance of a relative clause related to its embedding position in a main clause?
2) What is the relationship between the length of a relative clause and the dependency distance?
3) What is the relationship between the length of a relative clause and the dependency direction?
4) What is the regularity of dependency direction of a relative clause?
5) What are the influences of a relative clause on the dependency direction and dependency distance of its main clause?

The above questions are intended to find out the effects of the embedding position and the length of relative clauses on processing or comprehending difficulty of relative clauses and complex sentences in which relative clauses are embedded as well as the exploration of the universalities of different types of relative clauses. The research findings will help us to have a deeper understanding of syntactic difficulty of relative clauses and the corresponding complex sentences.

2 Methods and Materials

Under the framework of dependency grammar, the syntactic structure of a sentence is composed of the syntactic dependency relations, which are generally considered to have the following three core properties (Tesnière 1959; Mel'čuk 1988; Hudson 2007; Liu 2009):

(1) Dependency relation is a binary relation between two linguistic units;
(2) Dependency relation is usually asymmetrical, with one of the two units acting as the governor and the other as dependent;
(3) Dependency relation is labeled, which should be distinguished by explicitly labeling the arc linking the two units that form a dependency relation as shown in Fig.1.

The linear distance between the governor (also called "head") *learn* and the dependent, *English,* having a syntactic relation within a sentence, is defined as dependency distance (abbreviated as DD) referring to the number of intervening words. In other words, the term introduced by Hudson is defined as "the distance between words and their parents, measured in terms of intervening words" (Hudson 1995: 16).

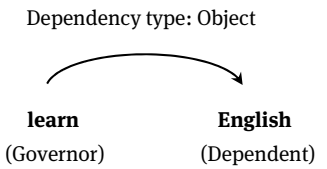


Fig. 1: Dependency relation of *learn English*

It must be pointed out that according to the calculation method proposed by Hudson, the dependency distance between *learn* and *English* is 0, which is unable to indicate dependency direction. Thus, in the present study, we used the calculation method proposed by Liu (2008: 163): for any dependency relation between the words W_a and W_b, if W_a is a governor (head) and W_b is its dependent, the dependency distance (DD) between the two words can be defined as the difference $a–b$. By this measure, the adjacent words have a DD of 1 (rather than 0 as is the case when DD is measured in terms of intervening words). According to the calculation method, we could distinguish dependency direction: if a is greater than b, the DD is a positive number, which means that the governor follows the dependent; if a is smaller than b, the DD is a negative number and the governor precedes the dependent. If dependency direction is not considered, for example, in a study measuring dependency distance, we can accordingly adopt, as a measure, the absolute value of dependency distance. In the present study, the dependency distance between *learn* and *English* is 1, and it is a head-initial dependency type. Generally speaking, a sentence includes more than one kind of dependency relation. Dependency distance of a sentence is the average value of absolute value of all the DDs and this average value is called mean dependency distance (MDD).

According to dependency parsing (Liu 2008; Hudson 2010), the processing of a word involves syntactically relating it to a previous word stored in memory. Therefore, as Lu et al. (2016: 2) pointed out, longer dependency distance means higher syntactic complexity and more intervening words may lead to more time-

based decay or more memory interference and render it more difficult to retrieve the syntactically related words. Liu et al. (2017: 171) also thought that dependency distance, measured by the linear distance between two syntactically related words in a sentence, is generally held as an important index of memory burden and an indicator of syntactic difficulty.

According to Liu (2008: 178), MDD can be used as a metric of syntactic complexity, reflecting the comprehending difficulty of a linear sequence of a sentence, and this has been verified in psycholinguistic experiments (Hudson 1996; Liu 2008).

From the psychological perspective, the comprehension of center-embedded structures has been found to be more difficult than that of right-branching sentences (Miller & Chomsky 1963; Weckerly & Elman 1992; Hudson 1995). The following sentence (1) with a center-embedded structure is more difficult to process than sentence (2) with a right-branching structure.

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

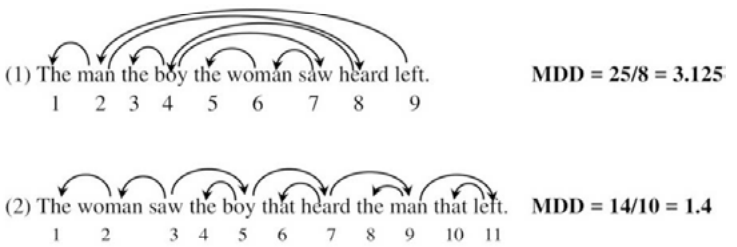


Fig. 2: Dependency structures and MDDs of (1) and (2)

As Fig. 2 shows, MDD of (1) is significantly longer than that of (2) (3.125 > 1.4), indicating that the former is more difficult to comprehend or process than the latter. The result is consistent with findings in psychological linguistics. The comparison suggests that MDD can be regarded as a syntactic complexity metric (Liu 2008: 169).

The value of DD refers to the linear distance between two linguistic units. The negative value stands for the fact that the head precedes the dependent (head-initial) while the positive value means that the dependent precedes the head (head-final).

On the basis of the three core properties, we can construct a syntactic dependency tree as a syntactic representation of a sentence or a sub-clause. In the present paper, directed acyclic graphs are chosen to illustrate dependency structure. In order to calculate MDD of all types of relative clauses, we have built the corresponding syntactic treebank, which can be used to train and evaluate a syntactic parser in computational linguistics (Abeillé 2003).

A sample complex sentence containing a relative clause is as follows:

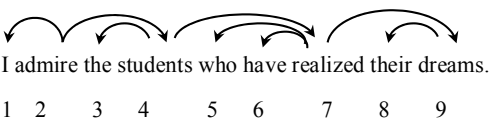


Fig. 3: Dependency structure of the sample sentence "*I admire the students who have realized their dreams.*"

For the present study, we have randomly selected one hundred complex sentences containing relative clauses, from *The New York Times* from January to April, 2016. To avoid the increase of more variables, the selected complex sentences and relative clauses meet the following requirements: each complex sentence includes only one type of sub-clause, relative clause and the subordinators of these relative clauses are not omitted. For example, in the complex sentence *There are some dreams that we desire to realize*, *that we desire to realize* is the relative clause, and *that* is the subordinator which is not omitted.

Relative clauses can be embedded in subject or object position of the main clause in a complex sentence. According to the embedding position of English relative clauses in main clauses, we have classified the one hundred relative clauses. We refer to the relative clause modifying subject as the relative clause embedded in subject (RCES) position and the relative clause modifying object as the relative clause embedded in object (RCEO) position. The relative clause, which is embedded in subject position of the main clause and whose antecedent functions as a subject in relative clause is marked as SS while the relative clause whose antecedent functions as an object in the main clause is marked as SO. If a relative clause is embedded in object position of the main clause and the antecedent functions as a subject in the relative clause, it is marked as OS. While the antecedent functions as an object in the relative clause, it is marked as OO. Based on previous studies, SS and OS are marked as SRC (Subject Relative Clause) while SO and OO are marked as ORC (Object Relative Clause). In addi-

tion to the above four types of relative clauses, the remaining relative clauses are regarded as Others.

Tab. 1 shows the distribution of the number of the one hundred relative clauses.

Tab. 1: Distribution of relative clauses

Subdivision of Relative Clauses		Number
RCES	SS	42
	SO	4
	Others	4
RCEO	OS	31
	OO	11
	Others	8
Total Number		100

Mean dependency distance of main clause (MDD_S) with n words can be calculated with the following equation in Equation (1).

$$MDDs = \frac{1}{n-1}\sum_{i=1}^{n-1}|DDi| \tag{1}$$

The mean dependency distance of English relative clause (MDD_c) with n words can be calculated with the following equation in Equation (2).

$$MDDc = \frac{1}{n}\sum_{i=1}^{n-1}(|DDi| + |DDmc|) \tag{2}$$

In Equation (1) and (2), n refers to the number of words in an English sub-clause or main clause; n–1 refers to the number of dependency relations; n also refers to the sum of n–1dependency relations in a sub-clause and the one dependency relation between main clause and sub-clause (n = [n–1]+1); DD_i is the dependency distance of the i-th dependency relation of the sub-clause; DD_{mc} is the dependency distance of the dependency relation between the sub-clause and its main clause in an English complex sentence.

Thus, for the sentence "*I admire the students who have realized their dreams*" in Fig. 3, the total number of dependency relations is n–1 = 9–1 = 8; the one dependency relation between main clause and sub-clause is the one between *students* and *realized*; the mean dependency distance of the relative clause, the MDD_c is [(2+1+2+1) + 3] / 5 = 1.8, as shown in Fig. 4.

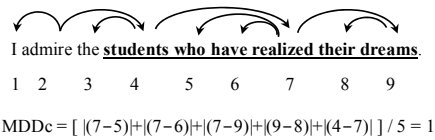


Fig. 4: Mean dependency distance of the relative clause

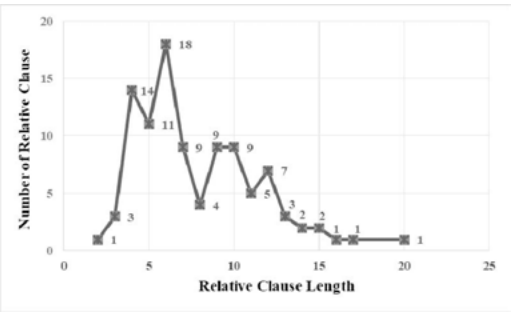


Fig. 5: Distribution of relative clause length

The treebank we have built consists of one hundred English complex sentences, with 1738 tokens and an average sentence length of 17.38 words. There exist the 100 relative clauses, with 784 tokens and an average sentence length of 7.84 words. The length distribution of these relative clauses is shown in Fig.5.

To explore the effect of the length of relative clause on the MDD of the relative clause, we separated the one hundred relative clauses into two sets (one set has 78 relative clauses with 2-10 words and the other has 22 relative clauses with 11-20 words) to clarify whether difference in relative clause length will lead to significant difference in their MDDs.

3 Results and Discussions

3.1 The probability distribution of dependency distance of relative clause

According to PDH (Kuno 1974) and SOHH (Hamilton 1994), processing difficulty of relative clauses may vary with the change of the embedding position. Dependency distance can be used as a metric to measure the syntactic complexity (Liu 2008: 172) and it can help us understand the differences of syntactic complexity of relative clauses better to investigate the probability distribution of dependency distance. Previous studies (Liu 2007; Ferrer-i-Cancho & Liu 2014; Wang & Liu 2017) also show that the distribution of dependency distance follow certain regularities in many languages, such as English and Chinese. Jiang and Liu (2015: 96) believe that the probability distribution of dependency distance of natural texts can contribute to our better understanding of their general features. Therefore, it is necessary for us to examine whether the probability distribution of dependency distance of relative clauses embedded in different positions, namely RCES and RCEO, follows certain regularities.

Liu (2007: 1) pointed out that the probability distribution of dependency distance can be better captured by right truncated Zeta distribution. In order to find out what kinds of models the probability distributions of DD of RCES and RCEO follow, we fitted our data to Altmann-Fitter (2013). The fitting results are presented in Tab. 2.

Tab. 2: Fitting of DD in RCES and RCEO to different distribution models

Types	Modified Zipf-Alekseev	Right truncated waring	Right truncated zeta
RCES R^2	0.9991	0.9964	0.9450
RCEO R^2	0.9990	0.9927	0.9399

Fig. 6 and Fig. 7 show the fitting results of a right truncated zeta to the DD of two types of relative clauses, RCES and RCEO.

The fitting results show that the DD of the two sets of relative clauses abide by the following probability distributions: Right truncated modified Zipf-Alekseev (a, b; n = x-max, a fixed), Right truncated Waring (b, n) and Right truncated zeta (a; R = x-max), as shown in Tab. 2.

The two fittings of right truncated waring are almost perfect, with the R^2 value of RCES being 0.9964 and that of RCEO being 0.9927. Meanwhile, the fittings of right truncated zeta are not as good as others, but they are also acceptable (R^2 of RCES = 0.9450 and R^2 of RCEO = 0.9399).

Tab. 2 indicates very similar probability distributions of the DD of the two types of relative clauses, which seems insensitive to their embedding positions. In addition, Fig. 6 and Fig. 7 suggest that, in the relative clauses, the adjacent dependency relation and dependency relation whose dependency distance is 2 are dominant. These findings imply that the users of relative clauses tend to minimize dependency distance, no matter which position of the main clause they are embedded in. This may be attributed to human working memory limitations. When dependency distance of a sentence or structure is shorter, it requires less effort to process it (Gibson 2000; Wang & Liu 2017).

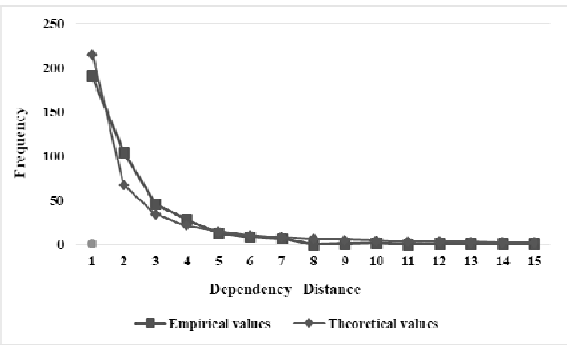


Fig. 6: Fitting of a right truncated zeta to the dependency distance of RCES

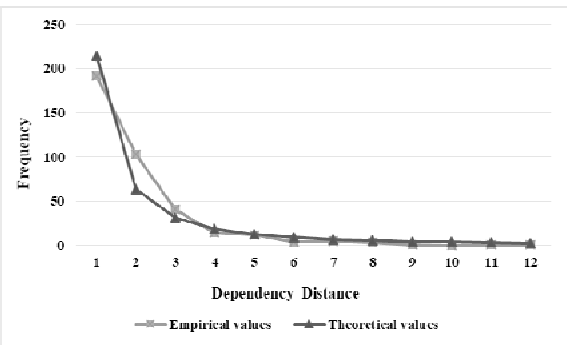


Fig. 7: Fitting of a right truncated zeta to the dependency distance of RCEO

3.2 The relationship between relative clause length and dependency distance

As for the effect of length of relative clause on attachment preference of relative clause in the clause, in a written questionnaire experiment, Hemforth et al. (2015: 43) tested German, English and Spanish and found that in all three languages, more high attachment interpretations were observed for long relative clauses than for short ones. Fodor (1998) proposed the Balanced Sister Hypothesis and held that a short relative clause (e.g. who drank) might preferably attach low position to achieve balance. These demonstrate that the length of relative clauses does influence their attachment position in main clause, even though there has not been any research on the effect of different positions of relative clauses.

According to PDH (Kuno 1974) and SOHH (Hamilton 1994), OS or OO is easier to comprehend than SS or SO. However, our research result indicates that there exists significant difference in MDD between the set with 2-10 words and the other with 11-20 words (p-value = 0.002; the values are 1.90 and 2.26). Here, the position of relative clauses has not yet been taken into consideration, thus it is unknown how the embedding position influences the MDD of relative clauses.

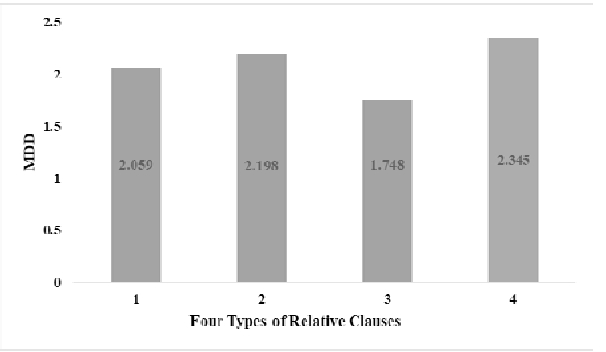


Fig. 8: Influences of relative clause length on MDD
Note: 1 = 38 relative clauses with 2-10 words of RCES; 2 = 12 relative clauses with 11-20 words of RCES; 3 = 40 relative clauses with 2-10 words of RCEO; 4 = 10 relative clauses with 11-20 words of RCEO

To verify the relationship between MDD and embedding position, we have analyzed the differences of MDD of RCES (38 relative clauses with 2-10 words and 12 relative clauses with 11-20 words) and RCEO (40 relative clauses with 2-10 words

and 10 relative clauses with 11-20 words) with the changes of relative cause length respectively. The results suggest that there is no significant difference between the MDDs of two types of relative clauses of RCES (p-value = 0.456; the mean values are 2.06 and 2.18), but significant difference exists between the MDDs of two types of relative clauses of RCEO (p-value = 0.000; the mean values are 1.75 and 2.35), which indicates that, for RCEOs, the MDD may change with the change of relative clause length.

In addition, we also compared the MDD of the 10 relative clauses with 11-20 words and that of two types of relative clauses belonging to the SRC. The results show that there is no significant difference between them (p-value > 0.05), as presented in Fig. 8. PDH and SOHH all demonstrate that the position of a relative clause in the main clause is the key factor influencing comprehension difficulty, but our results indicate that besides embedding position, relative clause length shouldn't be ignored, because greater length of relative clauses may result in longer MDD and longer MDD makes it more difficult to comprehend the relative clauses.

In conclusion, the three hypotheses, PDH, NPAH and SOHH, may be more convincing if the influence of the relative clause length is taken into consideration.

3.3 The relationship between relative clause length and dependency direction

Dependency direction refers to the relative position of two linguistic units (head and dependent) that have a dependency relation in a sentence (Jiang & Liu 2015: 102). When the head follows the dependent, the dependency relation is head-final; when the dependent follows the head, the dependency relation is head-initial. So far, it is still unknown whether relative clause length has anything to do with dependency direction of the relative clause.

Dependency direction reflects the linear order of two linguistic components that have syntactic relationship and can better reflect the language structure in a sentence (Liu et al. 2009: 518). It can also help better distinguish between genres.

Dependency relations between adjacent words bring about the lowest cognitive load, which has been confirmed by previous studies (Ferrer-i-Cancho 2004; Liu 2008) and supported by related theoretical analysis (Ferrer-i-Cancho 2004 2013).

Based on twenty languages, Liu (2008: 181) found that almost 50% of the dependency relations are adjacent and the number of the adjacent dependen-

cies is one of the important factors which influences mean dependency distance of a sentence. In addition, Collins (1996: 187) demonstrates that English is a language with rather fixed word order and finds that the percentage of adjacent dependency links is 74.2% for English language and meanwhile, only 12.1% of dependency links are formed at a distance of 2. The percentage of adjacent dependencies is 78% in Eppler's study (2005), and 61.7% in Jiang and Liu's (2015) study. According to Jiang and Liu (2015: 103), dependency distance is influenced by cognitive capacity, and it is because of the principle of least effort that adjacent dependencies account for more than 50% of dependencies. In our dependency syntactic treebank, the adjacent dependency links of English relative clauses account for 48.98%, which is close to the average of human languages. It needs to be emphasized that the above different percentages with regard to adjacent dependency links might result from the differences of research data or annotation scheme.

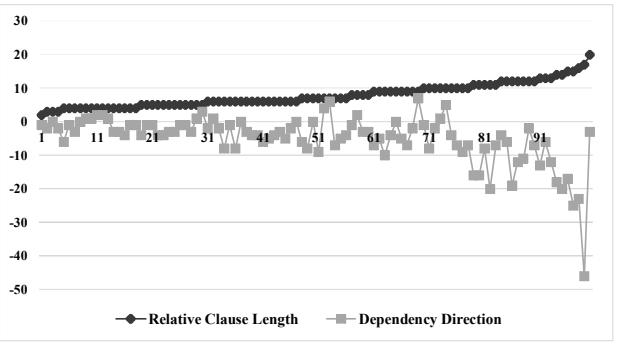


Fig. 9: Distribution of dependency direction with relative clause length

We could infer from Fig. 9 that there exists a significant relationship between the length of relative clauses and dependency direction (Pearson correlation = −0.640, *p*-value = 0.000). It is obvious that when the length of relative clauses is equal to and longer than 10, the value of DD becomes negative, which indicates that dependency relations of relative clauses are head-initial. In other words, 10 (the length of relative clauses) might be a key cut-off point, which determines the relationship between the length of relative clauses and dependency direction. To confirm this point, we divided one hundred relative clauses into two sets, according to their length (one being shorter than 10, including 69 clauses and the other equal to and longer than 10, including 31 clauses) and conducted correlation analysis. The results show that when the length of relative clauses is

shorter than 10, there is no significant correlation between the length and dependency direction (Pearson correlation = −0.228, *p*-value = 0.059). When the length of relative clauses is longer than 10, there exists significant correlation between them (Pearson correlation = −0.523, *p*-value = 0.003). It is obvious that there exists a distinct tendency that the longer the relative clause is, especially when the length is equal to and longer than 10, the more likely the relative clause is to possess the dependency relation of head-initial.

3.4 Distribution of dependency direction of relative clauses

The value of dependency distance can be either positive or negative, which signifies the dependency direction and also shows the linear order of governor (head) and dependent (Greenberg 1963; Dryer 1992). In terms of dependency direction, some languages seem to be consistently head-final and some others are head-initial (Hudson 2003).

Perceptual Difficulty Hypothesis (Kuno 1974) holds that due to the limitations of short-term memory, the relative clause embedded in subject position is more difficult to process than the one embedded in object position, which suggests that the position is a key factor influencing the processing difficulty of relative clauses. However, it is unknown whether dependency direction of a relative clause is relevant to its embedding position in the main clause. So we calculated the distribution of DD of relative clauses to find out whether the embedding position of a relative clause influences its dependency direction.

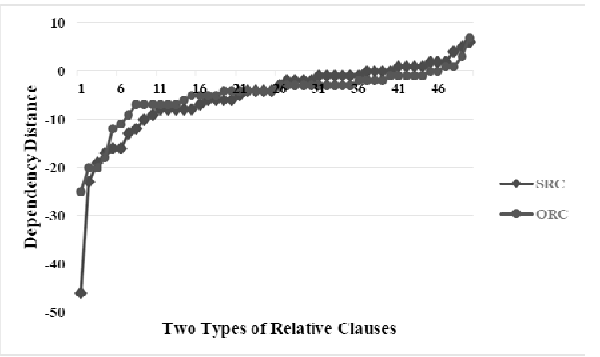


Fig. 10: Distributions of dependency direction

In the 50 RCESs, the dependencies within 36 RCs present an overall tendency to be head-initial (mean value of dependency distance is −8.06) and the dependency directions of 10 relative clauses are head-final (mean value of dependency distance is 2.50), while in the 50 RCEOs, the dependencies within 44 RCs present an overall tendency to be head-initial (mean value of dependency distance is −5.77) while the dependency directions of 4 relative clauses are head-final (mean value of dependency distance is 3.00). As for the dependency distances of RCESs and RCEOs, statistical test indicates that *p*-value is 0.754, which means the distributions of dependency direction within RCESs and RCEOs are similar. In addition, Fig. 10 also suggests that for both types of relative clauses (RCES and RCEO), their internal dependencies generally present similar dependency directions , which indicates that the dependency directions of relative clauses have nothing to do with their embedding positions in main clauses. In addition, we could infer from the above figures that the values of DD are largely negative, indicating that RCES and RCEO are more head-initial (*p*-value = 0.003, mean values are 1.85 and −6.80).

3.5 Effects of relative clauses on dependency direction of main clauses

Relative clauses are always embedded in the main clauses to form complex sentences. Then, what syntactic influences may relative clauses have on the main clauses? Are there corresponding changes of dependency directions? There are scarcely any discussions in the three previous hypotheses on relative clauses (Kuno 1974; Keenan & Comrie 1977; Hamilton 1994).

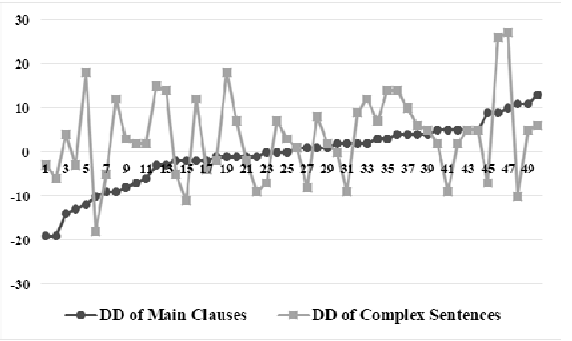


Fig. 11: Changes of dependency direction of main clauses of RCES

We calculated the DDs of main clauses in which 50 RCESs are embedded and the corresponding complex sentences, as shown in Fig. 11. Because of the embedding of RCESs, DDs of almost all main clauses change and DDs of main clauses of 34% RCESs become longer than before. The dependency direction of the corresponding complex sentences is still head-final; DDs of main clauses of 62% RCESs become shorter than before. Meanwhile, the dependency of the corresponding complex sentences is still head-initial.

From the above analysis and Fig. 11, we can infer that as for the RCESs, the embedded relative clauses make the dependency directions of their main clauses become more head-initial (p-value = 0.035).

In addition, we also calculated the DDs of main clauses where 50 RCEOs are embedded and the corresponding complex sentences, as presented in Fig. 12. Because of the embedding of RCEOs, DDs of almost all main clauses change and DDs of main clauses of 58% RCEOs become longer than before and the dependency relations of the corresponding complex sentences are still head-initial; DDs of main clauses of 36% RCEOs become shorter than before and the dependency relations of the corresponding complex sentences are still head-final.

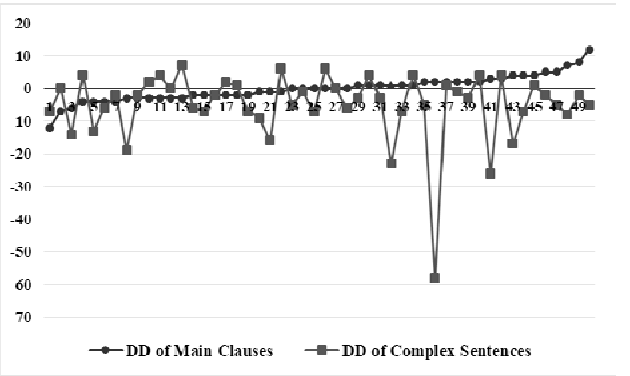


Fig. 12: Changes of dependency direction of main clauses of RCEO

However, the RCEOs don't lead to significant changes in the dependency directions of their main clauses (p-value = 0.058), as can be seen in Fig.12. That is to say, the dependency directions of the corresponding complex sentences are mixed. Liu (2010: 1570) proposed that most languages favor a balanced distribution of the dependency direction and the percentages of head-initial and head-final dependency relations are 48.8% and 51.2% respectively in English, which

means the dependency relations of English structures tend to be mixed. This is nearly consistent with our research result in reference to the RCEO.

The present research results show that the changes of dependency directions from main clauses to the corresponding complex sentences are influenced by the embedding position of relative clauses. When relative clauses are embedded in the object position of main clauses, the changes of dependency directions are more obvious.

3.6 Effects of relative clauses on dependency distances of main clauses

Then, what influence will relative clauses have on the MDDs of main clauses? To this question, this section will attempt an answer.

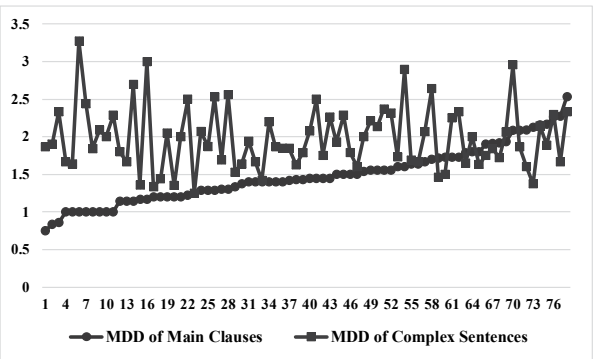


Fig. 13: Changes of MDD from main clauses to complex sentences

Fig. 13 presents the distributions of MDDs of main clauses and complex sentences. We can see from it that MDDs of the vast majority of main clauses become significantly longer. The embedding of relative clauses increases the MDD of 83% of main clauses (p-value = 0.000), but reduces the MDD of only 16% of main clauses (p-value = 0.002). In addition, we have statistically compared mean values of the increased distance and the reduced distance. The result shows that the increased distance is significantly greater than the reduced distance (p-value = 0.002, the mean values are 0.71 and 0.29).

Based on the above statistical analyses, we conclude that the embedded relative clauses increase the mean dependency distance of their main clauses significantly.

Hiranuma (1999: 316) reported that MDD of spoken English is 1.386 while Liu (2008: 174) found that MDD of English news is 2.543. In our recent research, MDD of English complex sentences selected from *The New York Times* is 2.039, which is different from Hiranuma's (1999) average value, but very close to Liu's (2008). In addition, Oya (2013: 47) calculated 1,196 sentences of English news and found that their average dependency distance is 3.34, which seems to be significantly longer than the result in the current study.

4 Conclusions and Implications

The research results based on texts selected from *The New York Times* indicate that mean dependency distance of English relative clauses embedded in the position of object of main clauses may change with the change of the length of relative clauses, especially for those with 11-20 words. We are unsure of the reason for this phenomenon, and that is probably worthy of future study based on the larger samples. However, mean dependency distance of relative clauses embedded in the position of subject has nothing to do with the length of relative clauses. According to PDH (Kuno 1974), the relative clause embedded in subject position of main clause is more difficult to process. The main cause for this hypothesis lies in the limitations of human short-term memory, because the relative clause embedded in subject position separates subject of main clauses and other components, and it is difficult to establish syntactic relation among separated components in our memory. In addition, longer MDD makes it more difficult to comprehend a sentence. Given these two factors, we think it makes sense that MDDs of relative clauses embedded in object position of main clauses don't vary with the change of the length of relative clauses. Our study, which takes the length of relative clauses into consideration and provides consistent empirical evidence that MDDs of relative clauses are related to their length, makes a complement to PDH.

The probability distributions of the DDs of relative clauses are insensitive to their embedding positions. The dependency distance distributions of two sets of relative clauses, 50 RCESs and 50 RCEOs, follow similar distribution models — a fact indicating that the distributions of dependency distance have regularities. The root might lie in that human languages have a tendency to minimize the

dependency distance, which is largely realized through a high percentage of adjacent dependencies.

In addition, as for the relationship between dependency direction and relative clause length, the research shows that the longer the length of relative clause is, the more likely the relative clause is to possess head-initial dependency relations. Dependency directions of relative clauses aren't influenced by embedding positions and dependency directions of relative clauses are more likely to be head-initial. The changes of dependency directions of main clauses are influenced significantly by the embedding position of relative clauses.

What is more, relative clauses significantly influence the MDDs of complex sentences. The MDD of 83% of main clauses is significantly longer after the attachment of a relative clause, while the MDD of only 16% of main clauses decreases after such an attachment. The present study demonstrates the differences and distributions of dependency distance and dependency directions of relative clauses and also exhibits their effects on main clauses.

However, there is still large room for improvement. The research data include one hundred relative clauses, whose relative pronouns or adverbs aren't omitted. A larger sample with more diverse relative clauses may provide us with new findings regarding the effects of relative clauses on main clauses and distributions of dependency directions of English relative clauses. Methodologically, if we separated the different sets in terms of the length of relative clauses, such as 1-5 words and 6-10 words, a more accurate picture may be drawn to reveal the relationship between the length of relative clauses and their dependency directions.

References

Abeillé, Anne. 2003. *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer Academic Publishers.

Altmann-Fitter. 2013. Altmann-fitter User Guide. The Third Version. Downloadable at: https://www.ram-verlag.eu/software-neu/software/.

Bates, Elizabeth. 1999. Processing complex sentences: A cross-linguistic study. *Language and Cognitive Processes, 14* (1), 69–123.

Collins, Michael John. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, 184–191.

Diessel, Holger & Michael Tomasello. 2005. A new look at the acquisition of relative clauses. *Language, 81* (4), 882–906.

Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language, 68* (1), 81–138.

Eppler, Eva Maria. 2005. The syntax of German-English code-switching. Ph. D. Dissertation. London: University College London.
Ferrer-i-Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical Review E, 70*, 056135.
Ferrer-i-Cancho, Ramon. 2013. Hubiness, length and crossings and their relationships in dependency trees. *Glottometrics, 25*, 1–21.
Ferrer-i-Cancho, Ramon & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory, 5* (2), 143–155.
Fodor, Janet Dean. 1998. Learning to parse? *Journal of Psycholinguistic Research, 27* (2), 285–319.
Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68* (1), 1–76.
Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Edward Gibson, Alec Marantz, Yasushi Miyashita & Wayne O'Neil (Eds.), *Image, Language, Brain* (pp. 95–126). Cambridge, MA: The MIT Press.
Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (Ed.), *Universals of Language* (pp. 73–113). Cambridge, MA: The MIT Press.
Hamilton, Robert L. 1994. Is implicational generalization unidirectional and maximal? Evidence from relativization instruction in a second language. *Language Learning*, *44* (1), 123–157.
Hawkins, John A. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.
Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.
Hemforth, Barbara, Susana Fernandez, Charles Jr. Clifton, Lyn Frazier, Lars Konieczny & Michael Walter. 2015. Relative clause attachment in German, English, Spanish and French: Effects of position and length. *Lingua, 166*, 43–64.
Hiranuma, So. 1999. Syntactic difficulty in English and Japanese: A textual study. *UCL Working Papers in Linguistics, 11*, 309–322.
Hou, Jiandong. 2011. An OT analysis of the difficulty hierarchy affected by accessibility and animacy in the acquisition of English relative clauses by Chinese EFL learners. *Foreign Language Teaching and Research, 43* (5), 702–711.
Hudson, Richard. 1995. *Calculating Syntactic Difficulty*. Unpublished paper.
Hudson, Richard. 1996. The difficulty of (so-called) self-embedded structures. *UCL Working Papers in Linguistics, 8*, 283–314.
Hudson, Richard. 2003. The psychological reality of syntactic dependency relations. In *MTT 2003*, Paris.
Hudson, Richard. 2007. *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
Hudson, Richard. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.
Izumi, Shinichi. 2003. Processing difficulty in comprehension and production of relative clause by learners of English as a second language. *Language Learning, 53* (2), 285–323.
Jiang, Jingyang & 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.

Keenan, Edward L. & Bernard Comrie. 1977. Noun phrase accessibility and universal Grammar. *Linguistic Inquiry, 8* (1), 63–99.
Kuno, Susumu. 1974. The position of relative clauses and conjunctions. *Linguistic Inquiry, 5* (1), 117–136.
Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics, 15*, 1–12.
Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9* (2), 159–191.
Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.
Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120* (6), 1567–1578.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.
Liu, Haitao, Yiyi Zhao & Wenwen Li. 2009. Chinese syntactic and typological properties based on dependency syntactic treebanks. *Poznań Studies in Contemporary Linguistics, 45* (4), 509–523.
Lu, Qian, Chunshan Xu & Haitao Liu. 2016. Can chunking reduce syntactic complexity of natural languages? *Complexity, 21* (s2), 33–41.
Marefat, Hamideh & Ramin Rahmany. 2009. Acquisition of English relative clauses by Persian EFL learners. *Journal of Language and Linguistic Studies, 5* (2), 21–48.
Mel'čuk, Igor′ Aleksandrovič. 1988. *Dependency Syntax: Theory and Practice*. Albany, NY: State University Press of New York.
Miller, George A. & Noam Chomsky. 1963. Finitary models of language users. In Luce Robert Duncan, Robert Bush & Galanter Eugene (Eds.), *Handbook of Mathematical Psychology, Vol. 2* (pp. 559–588). New York: Wiley.
Oya, Masanori. 2013. Degree centralities, closeness centralities, and dependency distances of different genres of texts. In *Selected Papers from the 17th international conference of Pan-Pacific of Applied Linguistics* (pp. 42–53).
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
Reali, Florencia & Morten H. Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language, 57* (1), 1–23.
Romaine, Suzanne. 1984. Relative clauses in child language, pidgins, and creoles. *Australian Journal of Linguistics, 4* (2), 257 - 281.
Street, James A. 2017. This is the native speaker that the non-native speaker outperformed: Individual, education-related differences in the processing and interpretation of Object Relative Clauses by native and non-native speakers of English. *Language Sciences, 59*, 192–203.
Tesnière, Lucien. 1959. *Eléments de Syntaxe Structural*. Paris: Klincksieck.
Wang, Yaqin & Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences, 59*, 135–147.
Warren, Tessa & Edward Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition, 85* (1), 79–112.
Weckerly, Jill & Jeffrey L. Elman. 1992. A PDP approach to processing center-embedded sentences. In William Schmidt & Thomas Shultz (Eds.), *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 72–77). Hillsdale, NJ: Erlbaum.

Xiao, Yunnan & Jie Lv. 2005. Chinese learners' interlanguage development in the acquisition of several types of English relative clauses. *Foreign Language Teaching and Research, 37* (4), 259–264.

Chunshan Xu

Differences between English Subject Post-modifiers and Object Post-modifiers: From the Perspective of Dependency Distance

Abstract: Two hypotheses are put forward in this study. The first is that, constrained by the tendency toward dependency distance minimization, the post-modifiers of subjects probably tend to be shorter than those of objects, and the second is that, to reduce the processing difficulty invoked by long dependency distance, punctuation marks might be used more frequently after subject post-modifiers, especially the long ones, than after object post-modifiers. To test these two hypotheses, 320 subjects and 320 objects in SVO or SVC structures are collected to investigate their modifiers, and the results largely support these two hypotheses.

Keywords: dependency distance; subject; object; post-modifier

1 Introduction

The words of a sentence are linearly presented, one after another, which lead to the concept of dependency distance, defined as the number of words intervening between two syntactically related words(Hudson 2010), as shown in Fig. 1, where a long-distance dependency can be found between the predicate verb “hated” and the subject “man”.

Fig. 1: Dependencies and dependency distance in a sentence

Chunshan Xu, Anhui Jianzhu University, Hefei, P.R.China, adinxu@126.com

https://doi.org/10.1515/9783110573565-013

Due to the linear presentation of sentences, language comprehension is often held to be carried out word by word, in terms of both phrase structure grammar (PSG) (Kay 1980; Hawkins 2004) and dependency grammar (DG) (Tesnière 1959; Hudson 2007; Liu 2009), and it is often argued that processing difficulty may be quantified by the dependency distance, which lead to the so-called locality effect, bearing on the limited working memory of human being (Alexopoulou & Keller 2007; Gibson 1998, 2000; Hawkins 1994, 2004; Hudson 2007, 2010; Liu 2008). In other words, the more intervening words between two syntactically related words, the heavier load on working memory, and thus the more difficult to establish syntactic relation between them. Researches based on real language corpora have repeatedly reported a tendency in various languages to minimize dependency distance (Hawkins 2004; Ferrer-i-Cancho 2006; Liu 2007a, 2007b, 2008; Temperley 2007, 2008), a tendency which is held as resulting from limited working memory of human beings, shaped by the principle of least effort (Zipf 1949).

In natural languages, the tendency toward dependency distance minimization is largely realized through syntactic organizations, especially the regular patterns in word order. Many syntactic patterns, overt or covert, are moulded by the cognitive constraint of working memory to reduce dependency distance. These syntactic patterns or syntactic regularities may be apparent, as reflected by grammatical rules that are well-entrenched, familiar to all of us. But it is also possible that they may be obscure, hidden in syntactic phenomena, and can only be unveiled as statistical preferences through quantitative study of corpus. These syntactic patterns and regularities meet the requirement for DDM and thus give rise to the long-tail distribution of dependency distance in natural languages.

For example, the positions of modifiers of nouns are flexible in English, a SVO language, where both the subject and the object may take modifiers, before or after them. For SVO languages like English, subjects usually precede predicate verbs, with post-modifiers intervening between them and the predicate verbs and objects normally succeed predicate verbs, with pre-modifiers intervening between them and the predicate verbs. It then seems that modifiers of nouns are syntactically free in terms of positions, entirely up to the semantics. However, noun modifiers in English may present some obscure patterns in their positions, regularities that function to reduce dependency distance or the processing difficulty of dependency distance. This possibility is the main concern of this study. To answer these questions, we need to respectively investigate into the modifiers of subjects and objects. In English, the post-modifiers are usually much longer than pre-modifiers. Therefore, the influence on dependen-

cy distance largely comes from the post-modifiers, on which the present study focuses. In English, subject post-modifiers will increase the distance between subjects and the predicate verb to which they are connected, whereas object post-modifiers have no influence on the distance between the objects and the predicate verbs, as can be seen in Fig. 1. Hence, if DDM is a universal cognitive requirement of human languages, it might be expected that the use of post-modifiers is somewhat constrained by the pressure for DDM, and the post-modifiers of subjects probably tend to be shorter than those of objects, which makes the first hypothesis of this study.

Despite the overall tendency toward dependency distance minimization, long-distance dependencies are unavoidable in natural languages (Liang et al. 2017; Liu et al. 2017; Futrell et al. 2015; Liu 2007a; Gildea & Temperley 2010). Previous corpus-based studies (Futrell et al. 2015; Liu 2007a; Gildea & Temperley 2010) have indicated a fact that no language minimizes its MDD to the theoretical minimum. The reason may be that, in addition to the cognitive constraint of limited working memory, language systems should meet other needs, such as predictability maximization, pragmatic function, precise and reliable communication, discourse coherence, etc. These needs may give rise to sporadic long dependencies in natural languages. If DDM was the sole requirement, no subject would take post-modifiers, which is obviously not the case: other communication needs often render it inevitable to use post-modifiers after subjects, though that may lead to long dependency distances between subjects and predicates.

However, long distance does not necessarily mean heavy processing load. As a complex system, language is capable of self-organization and self-adaptation. It may come up with, during its evolution, some strategies to offset the possible processing complexity caused by long-distance dependencies. For example, chunking is an important way to improve processing efficiency and reduce processing time. Therefore, proper chunking may reduce the difficulty of long distance dependencies (Christiansen & Chater 2016). Corpus-based studies of English seem to suggest a linguistic regularity that functional elements like complementizer "that" are more likely to be used in long dependencies than in short ones (Jaeger 2010). Another study of Chinese indicates that the use of Chinese particle "*er* (而)" is somewhat subject to the length and the complexity of the intervening elements (Xu 2015). These functional elements, with their very high frequency, may serve as conspicuous chunk markers, facilitating the processing of intervening elements without neglecting important key constituents. Similar strategies may well be expected to be found in long English subject-verb relations, where punctuation marks like comma may have similar

functions: they might overtly segment the long structure into chunks to facilitate language processing. So the second hypothesis of this study is that punctuation marks might be used more frequently after subject post-modifiers, especially the long ones, than after object post-modifiers.

2 Materials and Results

The research materials of this study are taken from two English textbooks for English majors (Contemporary College English, Volume 5 and Volume 6[1]). From each volume, we selected eight articles, including speeches, essays, short stories and proses. From each article, we chose 40 subjects and 40 objects, which make 320 subjects and 320 objects in total, to study their modifiers (both pre- and post-modifiers). In this study, we consider only the basic sentence patterns of SVO and SVC, also the two most commonly used ones. Other sentence patterns like SVOC (subject+verb+object+complement) and SVoO (subject+verb+indirect object+direct object) are not included in the present study, because the additional dependencies between V and C or O may interfere. For each subject and object, we manually measured the length of their modifiers and recorded the use of punctuation. Tab. 1 gives the mean lengths of subject and object modifiers.

Tab. 1: The mean lengths of subject and object modifiers

	Mean length of pre-modifiers	Mean length of post-modifiers
subject	1.13	2.35
object	1.25	4.23

As can be seen from Tab. 1, the pre-modifiers of both objects and subjects are rather short, with mean length slightly longer than 1, and there was no significant length difference between the pre-modifiers of subjects and objects (p = 0.07). This finding suggests that pre-modifiers have rather limited influence on dependency distance. However, significant difference was found between the mean lengths of subject post-modifiers and object post-modifiers (p = 0.001). In

1 Contemporary College English (Intensive Reading), Beijing: Foreign Language Teaching and Research Press, 2002.

other words, this finding seems to suggest that people tend to use longer post-modifiers after English objects than after subjects, which may be a piece of supportive evidence for the first hypothesis.

Then we further investigated into the frequencies of subject and object modifiers with different lengths. As can be seen in Fig. 2, the subject post-modifiers whose length is 0 or 1 are more frequent than object post-modifiers with the same length. However, when the length of post-modifiers exceeds 1, the frequencies of object post-modifiers are generally higher than those of subject post-modifiers. In brief, short subject post-modifier are more frequent than short object post-modifiers, while long subject post-modifiers are less frequent than long object post-modifiers. These differences may be responsible for above difference in the mean length.

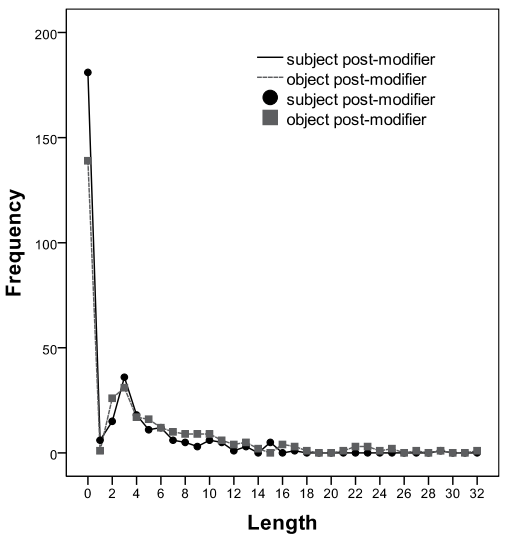


Fig. 2: Frequencies of the lengths of subject and object post-modifiers

Then, the Altmann Fitter[2] was applied to find out which specific distributions could match the frequencies of subject and object post-modifiers with different lengths. Probably owing to the small samples of this research, no distribution is

2 The Altmann Fitter is a programme for the fitting of univariate discrete probability distributions to frequency data, available at https://www.ram-verlag.eu/software-neu/software/.

found to acceptably match the raw data. Then the data was grouped into 11 sets at length interval of 3, but there was still no matched distribution. So the frequency data was regrouped into 7 sets at length interval of 5. This time, it was found that the frequencies of subject post-modifiers with different lengths may be captured by Mixed Poisson distribution (a=1.4602, b=0.3435, α=0.0098, p=0.5531), as can be seen in Fig. 3, while the frequencies of object post-modifiers seem to abide by both Mixed Poisson distribution (a=2.5071, b=0.4118, α=0.1043, p=0.598) and Right truncated Zeta distribution (a=2.0882, R=7.0000, p=0.6055), as can be seen in Fig. 4.

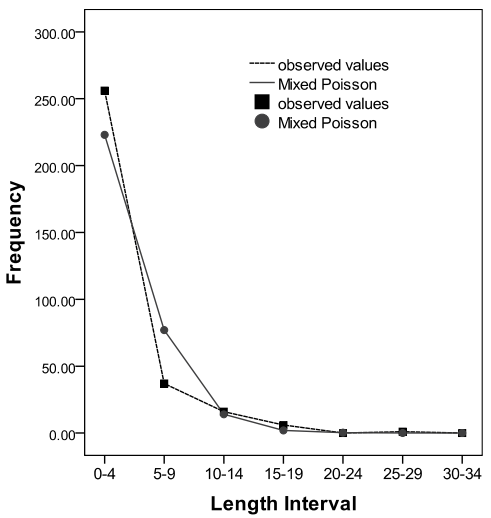


Fig. 3: Frequency distributions of subject post-modifiers

These findings seem to suggest different distribution patterns for object and subject post-modifiers, indicating that object post-modifiers tend to be shorter than subject post-modifiers, which may provide the evidence supporting the first hypothesis. The pressure for dependency distance minimization may have shaped the length difference between subject post-modifiers and object post-modifiers: for the length of former decides the distance of the key dependency between subject and predicate verb and should be minimized, while the length of latter has no significant effect on dependency distance. However, there is possibly another reason for these findings: they may be the results of, not only the dependency distance minimization, but the information structure of human languages, which typically presents a pattern of old information preceding new

information. Given this information structure, it may be expected that subjects, often occurring at the initial position of SVO or SVC structures, typically convey old, known information, requiring no additional specification. In this case, modifiers are usually unnecessary, and very often a pronoun suffices. In many languages, subjects are often omitted, if they are contextually known.

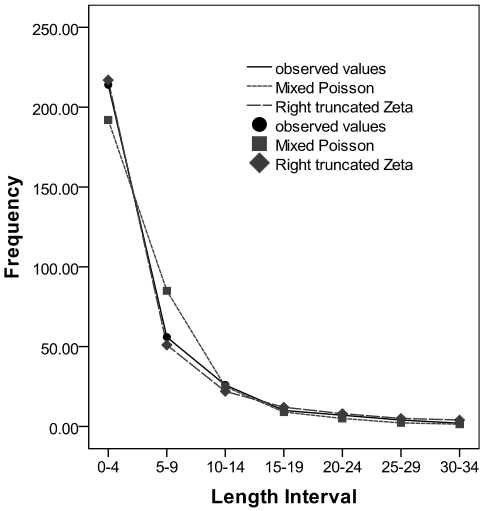


Fig. 4: Frequency distribution of object post-modifiers

On the contrary, the objects, often occurring at the end of a structure, very probably convey new information, requiring further specification, with modifiers more likely to appear. This difference is in agreement with the principle of iconicity (Haiman 1985): objects usually have more information content than subjects and therefore often carry longer subordinate elements. In view of this possibility, the above findings may be interpreted as the results of both DDM and the information-structure. Hence we need to minimize the influence of information structure to see whether the above differences still persist. So we screened the data to exclude all the pronouns, which are definitely known messages, and the subjects with less than 2 modifiers, which are very likely to be known messages, taking into account only those subjects and objects with at least 2 modifiers, which may well contain new messages. Then we recalculated the mean lengths of their modifiers to see whether the difference in mean length persists. Tab. 2 shows the results, which are similar to those given in Tab. 1.

Tab. 2: The mean lengths of subject and object modifiers (after screening)

	Mean length of pre-modifiers	Mean length of post-modifiers
subjects	1.540	4.100
objects	1.580	5.900
p-value	0.380	0.006

Similar to Tab. 1, Tab. 2 shows, again, that the mean length of object post-modifiers in English is longer than that of subject post-modifiers, and that the mean length of subject pre-modifiers is very similar to that of object pre-modifiers: both being rather short, slightly longer than 1. Tab. 2 also indicates that there is no significant difference in the means length between subject pre-modifiers and object pre-modifiers (p = 0.38). But significant difference has been found between the mean length of the post-modifiers of subjects and that of objects (p = 0.006). In short, the findings presented in Tab. 1 have almost been replicated in Tab. 2, which indicates a persistent tendency to use longer post-modifiers after English objects than after subjects, a tendency very likely to stem from the pressure for DDM, rather than from the information structure.

Then we further investigated into the frequencies of subject and object post-modifiers with different lengths. Fig. 5 shows that short subject post-modifiers seem to be more frequent than short object post-modifiers, but long subject post-modifiers tend to be less frequent than long object post-modifiers, roughly a similar finding to the previous one presented in Fig 2.

The Altmann Fitter was applied to the data again, to find out which distributions could match the frequencies of subject and object post-modifiers. Given the sparsity of sample, the data was grouped into 7 sets with a length interval of 5, and the fitting tests showed that the frequencies of subject post-modifiers may be captured by Mixed Poisson distribution (a=2.9097, b=0.2256, α=0.5373, p=0.0801) while the frequencies of object post-modifiers can be captured by Mixed negative binominal distribution (k=0.9717, $p1$=0.0042, $p2$=0.5324, α=0.0059, p=0.2065), as can be seen in Fig. 6 and Fig. 7. This finding suggests that, though the influence of information structure is reduced, subject and object post-modifiers seem to have different distribution patterns in length frequencies. This difference probably bears on the pressure for dependency distance minimization, not the information structure.

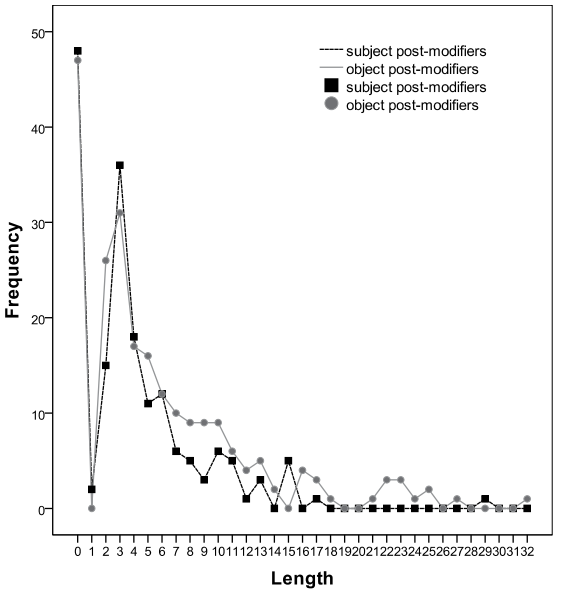


Fig. 5: Frequencies of subject and object post-modifiers (after screening)

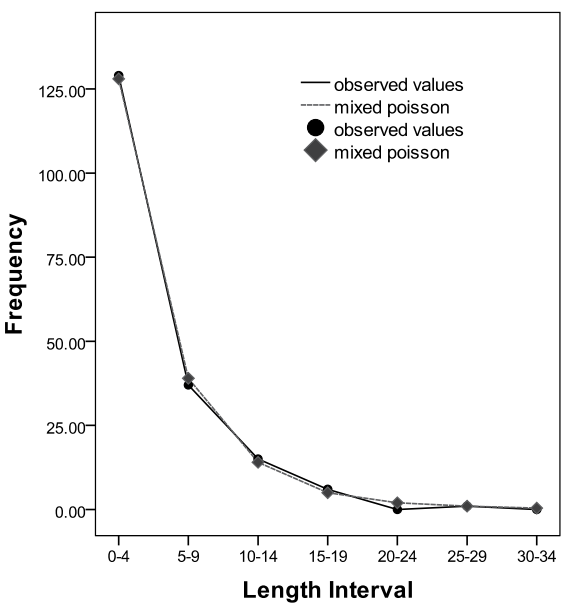


Fig. 6: Frequency distributions of subject post-modifiers (after screening)

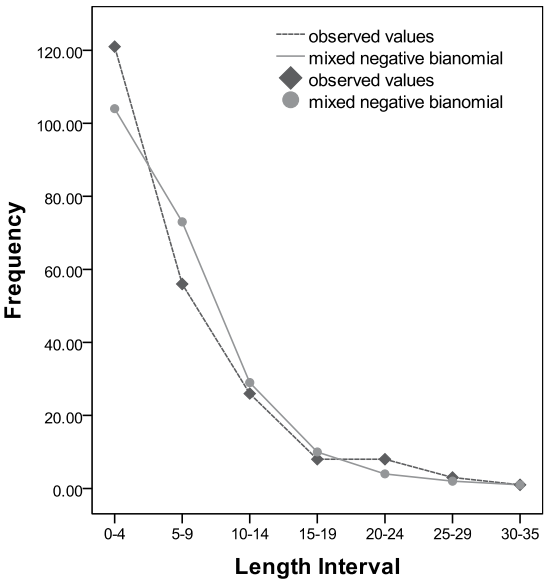


Fig. 7: Frequency distributions of object post-modifiers (after screening)

Dependency distance minimization might be a universal tendency, but it is undeniable that long dependencies are repeatedly found in natural languages, which may result from some communicative needs that temporarily override the pressure to reduce dependency distance, or rather, the memory load (Liu et al. 2017; Xu et al. 2017; Xu & Liu 2015). However, as a system capable of self-adaption, language may come up with certain strategies to reduce the difficulty invoked by long distance, including the use of function elements to facilitate chunking (Liu et al. 2017; Xu & Liu 2015). These strategies might also be utilized when post-modifiers are rather long. That is, punctuation marks might be used more frequently after post-modifiers of subjects than after those of objects, especially when the modifiers are long. To test this hypothesis, investigations were conducted into the use of punctuation marks, with the results shown in Tab. 3 and Tab. 4. As can be seen in Tab. 3, 27 out of 317 subject post-modifiers use punctuation mark, while only 10 out of 317 object post-modifiers use punctuation marks, which indicates a higher frequency of punctuation marks used after subject post-modifiers than after object post-modifiers: about 9% of subject post-modifiers are followed by punctuation marks, and only about 3% of object post-modifiers are followed by punctuation marks. This difference is statistically significant (p=0.0001).

Tab. 3: Subject and object post-modifiers with punctuation marks

	object post-modifiers	subject post-modifiers
The numbers	10（317）	27（317）
The proportions	3%	9%

What is more, there seems to be a relation between the length of post-modifiers and the use of punctuation marks. As can be seen in Tab. 4, when punctuation marks are used, the average length of subject post-modifiers is significantly longer than when punctuation marks are absent (p=0.000), and a similar case is found in object modifiers (p=0.000), which suggests that the length is probably one important factor determining the use of punctuation marks. Furthermore, Tab. 4 indicates that, when punctuation marks are used, the average length of subject post-modifiers is 7.12, while the average length of object post-modifiers is 15.29, which is about twice that of the former. This result indicates that the post-modifiers of subject seem to have much more urgent need to use punctuation marks than those of objects.

Tab. 4: The mean lengths of post-modifiers with and without punctuation marks

	With punctuation mark	Without punctuation mark
Subject post-modifiers	7.12	2.90
Object post-modifiers	15.29	3.97

In summary, it can be seen from Tab. 3 and Tab. 4 that, punctuation marks are used more frequently with subject post-modifiers and that the length of a modifier may bear on the use of punctuation marks.

3 Discussion

Language use is based on human brain and human cognition. Therefore, its operation and evolution, its patterns and regularities, are to a considerable degree driven by the basic mechanisms of the human brain, subject to multiple constraints of human cognition (Xu & Liu 2015, Liu 2014). According to dependency grammar (Liu 2009; Hudson 2010), the basic units of a sentence are words,

which are linked via dependency relations to form a complete hierarchical dependency tree pivoted on the main verb. This tree visualizes the underlying syntactic structure of a sentence. In terms of dependency grammar, the purpose of syntactic analysis is to specify the dependencies between all the words in a sentence, and finally build a dependency tree for the sentence. During this process, the linear distance between two interdependent words has a significant effect on the difficulty of parsing (Gibson 1998). This may be attributed to the psychological mechanisms of parsing, especially the limited working memory (Hudson 1995), as has been discussed in the Introduction. In short, due to the limited capacity of working memory, there is a tendency toward dependency distance minimization in natural languages, which plays an important role in shaping the syntactic patterns of natural languages, including the above differences in the subject and object post-modifiers.

These differences in length may be accounted for in the light of dependency distance minimization in natural languages (Liu 2008). English is a SVO language, which means that subject post-modifiers will have much influence on dependency distance because they intervene between subjects and predicate verbs, while object post-modifiers have no influence on the dependency distance because the post-modifiers are at the end of the structure. Frequent long post-modifiers after the subjects mean frequent long dependencies between subjects and predicate verbs. From a cognitive perspective, the longer the dependency is, the more severe the memory decay or interference is, and hence, the more difficult it is to establish syntactic or semantic relationships (Hudson 1995; Gibson 1998; Liu et al. 2017) between the governor and the dependent. In order to reduce the complexity of language processing, the subject post-modifiers in English may be expected to be shorter than object post-modifiers, so as to reduce the distance between subject and predicate verbs. In comparison, the object post-modifiers might be longer because their length has little to do with the dependency distance of the structure. So, the observed difference in mean length between subject and object post-modifiers may well be the result of the need to reduce dependency distance.

However, languages are subject to not merely dependency distance minimization: as a complex system, a language is usually constrained by many factors (Köhler 2005). As a result, it is inevitable to find many exceptions to dependency distance minimization. Subjects sometimes may be followed by long post-modifiers for the sake of, for instance, effective communication. The present study shows that approximately 20% of the subject post-modifiers contain 5 or more words, which leads to quite a few long dependencies between subjects and predicate verbs, increasing the complexity of language processing. To re-

duce the difficulty of processing, English language may adjust itself to these long dependencies, coming up with some strategies to handle them. In other words, long-distance dependencies may be accompanied by certain syntactic patterns that reduce the difficulty of processing long dependencies, as the result of the constraint of working memory and the principle of least effort (Xu 2015; Xu & Liu 2015, Liu et al. 2017). A previous study of Chinese particle "*er*" shows that when the preposition "*wei*" (for) is separated from its governing verb by a long prepositional object, the particle "*er*" is likely to be used before the verb, playing the role of a chunk marker, which can facilitate segmenting the string into syntactic chunks, and help readers promptly grasp the key components in quick skimming (Xu 2015). This strategy seems to also apply to the subject post-modifiers in English: our research shows that about 9% of the subject post-modifiers are followed by punctuation marks, while only 3% object post-modifiers are used following punctuation marks. In other words, the punctuation marks may also play the role of chunk markers to facilitate processing.

For long dependencies between subjects and predicate verbs, punctuation marks may signal conspicuously the subject, the modifier and the predicate verb, which contribute much to chunking and skimming, improving processing efficiency. Accordingly, it might be reasonable to infer that longer post-modifiers imply higher likelihood to use punctuation marks, which is confirmed in our research, suggesting a close relation between the use of punctuation marks like commas and the length of post-modifiers. When punctuation marks are used, according to the present study, the average length of subject post-modifiers is 7.12. However, when punctuation marks are not used, the average length is only 2.9. This great disparity strongly suggests a close relation between the use of punctuation marks and the distance between subjects and predicate verbs. Unlike the subject post-modifiers, the object post-modifiers have nothing to do with the dependency distance, and hence the need to use punctuation marks is relatively less urgent: only 3% of subject post-modifiers are used following punctuation marks. However, the effect of length is still striking: when punctuation marks are used, the average length of object post-modifiers is 15.29; when punctuation marks are not used, the average length is only 3.97.

It is worth noting that there is still a significant portion of long subject post-modifiers that are used without punctuation marks. This may cause some difficulty in understanding the sentence. In fact, for many Chinese learners of English, too long a dependency between a subject and its predicate verb often constitutes an obstacle for quick and efficient reading, especially when punctuation marks are missing. Luckily, for the most part, the first word of a long post-modifier is functionally marked, be it a non-finite verb, a relative pronoun, or a

preposition, which may somewhat play the role of chunk markers, separating the subject from the modifier. In English learning, most students, after a period of training, can quickly rely on these structural markers to efficiently handle the long-distance subject-verb structure. In other words, for both Chinese and English, functional elements may play an important role in processing long-distance dependencies. In English, pre-modifiers of nouns are usually devoid of functional elements, which may bear on their short length.

However, punctuation marks probably are more conspicuous chunk markers than functionally marked element: the blank spaces they leave in the sentence are more eye-attracting, in accordance with the principle of iconicity. What is more, punctuation marks are often used in pairs, not only to separate the subject from the post-modifiers, but also to separate the post-modifiers from the predicate verb, which may help spot key elements in sentences and thus improve processing efficiency.

In fact, chunking may be one basic cognitive approach to language processing. For long structures, efficient chunking may be particularly important, whether or not the structures involve long-range dependencies. For instance, the present study has found that, when object post-modifiers are extremely long, punctuation marks like comma are used to intervene between the objects and the modifiers, though the length of post-modifiers is independent of dependency distance.

This study considers only SVO and SVC structures, excluding SVoO and SVOC structures. However, if dependency distance minimization is universally valid, it might be expected the indirect object (o) in the SVoO structures should take short post-modifiers, otherwise the dependency distance between the direct object (O) and the verb (V) may be too long, causing difficulties and obstacles to understanding. Similarly, the object (O) in SVOC may also be expected to be followed by short post-modifiers. Another issue that probably deserves our attention is the dependency distance of adverbials. Usually, the adverbials appear at either the initial or the terminal positions of sentences, and the dependency distance is therefore sensitive to the length of subject or object post-modifiers. That is, when there is an adverbial, the length of noun post-modifiers may be subject to certain restrictions. These possibilities are probably worthy of further investigations.

4 Conclusion

To conclude, the present study indicates that, for English SVO and SVC structures, subject post-modifiers generally tend to be shorter than object post-modifiers, which might bear on the pressure for dependency distance minimization. This study also indicates higher frequency of punctuation marks used after subject post-modifiers, especially the long ones. This pattern may be shaped by the need to reduce the processing difficulty of long dependency caused by subject post-modifiers. In brief, this study suggests that the general pressure of dependency distance minimization and the principle of least effort might have shaped some covert patterns in the post-modifiers in English SVO and SVC structures.

Acknowledgement: This work was partly supported by the National Social Science Foundation of China (Grant No. 18BYY015).

References

Alexopoulou, Theodora & Frank Keller. 2007. Locality, Cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, *83*(1), 110–160.

Christiansen, Morten & Nick Chater. 2016. The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences,* 39, e62.

Ferrer-i-Cancho, Ramon. 2006. Why do syntactic links not cross? *Europhysics Letters,* 1228–1235.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence for dependency length minimization in 37 languages. *PNAS*, *112*(33), 10336–10341.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita & Wayne O'Neil (Eds.), *Image, Language, Brain* (pp.95–126). Cambridge, MA: The MIT Press.

Gildea, Daniel & David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Sciences*, *34*(2), 286–310.

Haiman, John. 1985. *Natural Syntax: Iconicity and Erosion*. Cambridge: Cambridge University Press.

Hawkins, John. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, John. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hudson, Richard. 1995. *Measuring Syntactic Difficulty*. Unpublished paper. http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf.

Hudson, Richard. 2007. *Language Networks: The New Word Grammar.* Oxford: Oxford University Press.

Hudson, Richard. 2010. *An Introduction to Word Grammar*. Cambridge: Cambridge University Press.
Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.
Kay, Martin. 1980. Algorithm Schemata and Data Structure in Syntactic Processing. In *Proceedings of the Nobel Symposium on Text Processing*. Gothenburg, 1980.
Köhler, Reinhard. 2005. Synergetic linguistics. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (Eds.), *Quantitative Linguistics: An International Handbook* (pp.760–774). Berlin: Mouton de Gruyter.
Liang, Junying, Yuanyuan Fang, Qianxi Lv & Haitao Liu. 2017. Dependency distance differences across interpreting types: Implications for cognitive demand. *Frontiers in Psychology*, *8*, 2132.
Liu, Haitao. 2007a. Probability distribution of dependency distance. *Glottometrics*, *15*, 1–12.
Liu, Haitao. 2007b. Dependency relations and dependency distance: A statistical view based on Treebank. In *Proceedings of the 3rd International Conference on Meaning-Text Theory*, 269–278.
Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal Cognitive Science*, *9*(2), 159–191.
Liu, Haitao. 2009. *Dependency Grammar: From Theory to Practice*. Beijing: Science Press.
Liu, Haitao. 2014. Language is more a human-driven system than a semiotic system. Comment on Modeling language evolution: Examples and predictions. *Physics of Life Reviews*, *11*, 309–310.
Liu, Haitao, Chunshan Xu & Junying Liang. 2017. Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193.
Temperley, David. 2007. Minimization of dependency length in written English. *Cognition*, *105*(2), 300–333.
Temperley, David. 2008. Dependency length minimization in natural and artificial languages. *Journal of Quantitative Linguistics*, *15*(3), 256–282.
Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.
Xu, Chunshan. 2015. The use and the omission of Chinese conjunction “er”. *Journal of Shanxi University (Philosophy and Social Sciences Edition), 38*(2), 55–61.
Xu, Chunshan, Junying Liang & Haitao Liu. 2017. DDM at work: Reply to comments on “Dependency distance: A new perspective on syntactic patterns in natural languages”. *Physics of Life Reviews*, *21*, 233–240.
Xu, Chunshan & Haitao Liu. 2015. Can Familiarity lessen the effect of locality? A case study of Mandarin Chinese subjects and the following adverbials. *Poznań Studies in Contemporary Linguistics*, *51*(3), 463–485.
Zipf, George. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Hafner.

Xinying Chen*, Kim Gerdes

How Do Universal Dependencies Distinguish Language Groups?

Abstract: The present paper shows how the current Universal Dependency treebanks can be used for language typology studies and can reveal structural syntactic features of languages. Two methods, one existing method and one newly proposed method, based on dependency treebanks as typological measurements, are presented and tested in order to assess both the coherence of the underlying syntactic data and the validity of the methods themselves. The results show that both methods are valid for positioning a language in the typological continuum, although they probably reveal different typological features of languages.

Keywords: treebanks; Universal Dependencies; typology; language family; Directional Dependency Distance

1 Introduction

Measuring cross-linguistic similarities and differences is one of the primary tasks of modern typology (Bickel, 2007), both theoretically and empirically. Modern language typology research (Croft 2002; Song 2001), mostly based on Greenberg (1963), focuses less on lexical similarity and relies rather on various linguistics indices for language classification, and generally puts much emphasis on the syntactic order (word order) of some grammatical relations in a sentence (Haspelmath et al. 2005).

However, a similar idea was proposed by Tesnière, before Greenberg, which attempts to classify languages by using a full analysis of a sentence based on binary grammatical relations rather than putting attention on some basic relations such as OV or VO (Liu 2010). One intuition about the difference can be that some basic grammatical relations are sufficient for detecting cross-linguistic similarities and differences already, at least for that time, the other factors being correlated with the basic distinctions of type OV/VO. Yet, several studies sug-

Xinying Chen, Xi'an Jiaotong University, Xi'an, P.R.China; University of Ostrava, Ostrava, Czech Republic, chenxinying@mail.xjtu.edu.cn
Kim Gerdes, Sorbonne Nouvelle, Paris, France

https://doi.org/10.1515/9783110573565-014

gest that this is not empirically true and it seems that combined measures on all grammatical relations, not only on the verbal and nominal arguments, provide better typological indicators than one or several specific word order measures for language classification, which may lead to conflicting conclusions (Liu 2010; Liu & Li 2010; Abramov & Mehler 2011; Liu & Xu 2011, 2012; Liu & Cong 2013). An alternative idea would be that it is just too difficult to do the research based on a complete analysis right now due to the limitations of available resources. If this is the case, then it is important to systematically replicate or test previous studies with new available data, which is also one of the motivations of this study since a suitable, newly emerged, multi-language treebank data - the Universal Dependencies treebanks - has appeared.

New available data always attracts the interest of linguists. In the case of empirical typology study, Liu (2010) used dependency treebanks of 20 languages to test and discuss some principal ideas of Tesnière and the research provided very convincing results which are in favour of Tesnière's classification. Liu (2010) completed Tesnière's attempt by developing an innovative quantitative measure of word orders, namely, the proportions of positive and negative dependencies. This measurement captures well the differences between head-initial and head-final languages. Empirical data based on 20 languages is provided along with the measurements, which makes the discussion very convincing and indicates the potential of this measurement for new data sets. However, there is a significant problem concerning this kind of Treebank based studies. The differences of annotation schemes can distort the analysis results for similar language structures or phenomena. For instance, Liu (2010) had a problem of placing Danish appropriately in his language continuum due to the peculiar annotation scheme used for the Danish dependency treebank. Addressing this problem and testing the validity and effectiveness of these measures with new available data, the Universal Dependencies treebanks, is the second motivation of the present study.

The last question we want to address is whether there are new ways of quantifying word orders for language classification and how this could be done? Following the success of dependency distance in linguistic studies (Liu et al. 2017), we propose a new way of measuring word orders, namely, directional dependency distance for this task (see section 2 for more details). The rationale is that since dependency distance reflects some universal patterns of human languages, such as dependency distance minimization (Liu et al. 2017), the details of the distance distribution in treebanks should be valuable for comparing

the similarities and differences of languages. We test the proposed measurement with the Universal Dependencies treebanks. Results and discussions are presented below.

The paper is structured as follows. Section 2 describes the data set, the Universal Dependencies treebanks, and introduces the new approach, directional dependency distance. Section 3 presents the empirical results and discussions of applying Liu's proportional +/− dependency approach as well as the directional dependency distance approach to the Universal Dependencies treebanks for the language classification task. Finally, Section 4 presents our conclusions.

2 Materials and Methods

Following the idea of investigating the typological similarities and differences of languages based on authentic treebank data, the present work specifically focuses on whether and how the Universal Dependencies treebank set allows us to recognize language families based on purely empirical structural data.

Universal Dependencies (UD) is a project of developing a cross-linguistically consistent tree-bank annotation scheme for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective. The annotation scheme is based on an evolution of Stanford dependencies (de Marneffe et al., 2014), Google universal part-of-speech tags (Petrov et al. 2011), and the Interset interlingua for morphosyntactic tagsets (Zeman, 2008). The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary. UD is also an open resource which allows for easy replication and validation of the experiments (all Treebank data on its page is fully open and accessible to everyone). For the present paper, we used the UD_V2.0 dataset (Nivre et al., 2017) for our study since it was the most recent data when we started this work.

There are two notable advantages of using this data set for language classification studies. Firstly, it is the sheer size of the data set: It includes 70 treebanks of 50 languages, 63 of which have more than 10,000 tokens. Secondly, and most importantly, all UD treebanks use the same annotation scheme, at least in theory, as we shall see below. The few previous studies of empirical language classification based on treebank data (Liu 2010; Liu & Xu 2011, 2012) still had to rely on much fewer treebanks with heterogeneous annotation schemes. Although already relatively satisfying results were obtained, the question of

identifying the source of the observed differences remains unsolved: They could be actual structural differences between languages or simply annotation schema related differences (or even genre related differences). For instance, concerning the Danish problem mentioned in the previous section, UD can, to a certain extent, reduce the difficulty by providing a unified framework for all languages.

However, there are also drawbacks of the UD 2.0 scheme for this task. First, UD treebanks are of fundamentally different sizes. We remove the relatively sparse languages, namely treebanks with less than 10,000 tokens, from the dataset. General information about the treebanks used for this study is provided in Appendix 1 (taken from the Universal Dependencies project page, http://universaldependencies.org). Different treebanks of the same language are kept separate for consistency measures and then combined for the main classification tasks.

Secondly, some dependencies have arbitrarily been assigned to a left-to-right bouquet structure (all subsequent tokens depend on the first token). See Gerdes & Kahane (2016) for a description and for alternatives to this choice. Thereby, we only kept syntagmatic core relations and removed fixed, flat, conj, compound, and root relations from our measurements as their direction and length are universally fixed in the annotation guide and don't indicate any interesting difference between languages.

After preparing the data for the analysis, we applied the proportional approach (Liu 2010) to test the validity and effectiveness of this approach. To be specific, we computed the percentage of positive dependency relations (corresponding to a head-initial structure) versus the percentage of negative dependency relations (corresponding to a head-final structure) in a language.

Then, we applied our newly proposed approach, Directional Dependency Distance, for the same task. We define Directional Dependency Distance (DDD) as the product of the dependency distance and the direction, thus including negative values. More specifically, we computed the DDD per language by computing the difference of the node index and the governor index for each node, adding those values up and dividing by the number of links. The DDD of a language L is thus defined by its dependencies as follows:

$$\mathrm{DDD} = \frac{\sum distance(d)}{\sum frequency(d)} \tag{1}$$

To illustrate this point, we computed the Dependency Distance between the word *This* (word id 1) and the word *example* (word id 4) in the Fig. 1 as: 1 – 4 =

−3 (a head-final structure). The DDD of the sentence in Fig. 1 is: [(1 − 4) + (2 − 4) + (3 − 4)] / 3 = −2.

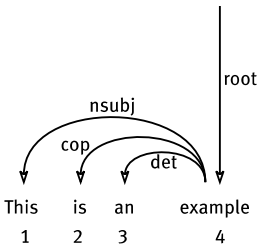


Fig. 1: A simple example of a UD tree

We then compare and discuss the results of these two approaches. Details are presented below.

3 Results and Discussion

In this section, we present the results of applying two dependency-based word order measurements to 63 treebanks which include 43 languages.

3.1 Proportional distribution of Universal Dependencies

The percentage of positive and negative dependency distances for each language is illustrated in Fig. 2 and corresponding values for treebanks are illustrated in Fig. 3. For detailed values see Appendix 2.

According to Fig. 2, this approach can distinguish Indo-European languages from other languages well. There are 12 languages that do not belong to the Indo-European families in our data set. Four of them, namely, Basque, Estonian, Finnish, and Hebrew (circled in Fig. 5), are placed into the Indo-European continuum where they should not appear. Simply put, we can say that the accuracy of distinguishing Indo-European languages of this approach is (43 − 4) * 100 / 43 = 90.7%, which seems like a rather good result. However, if we take a closer look at the Indo-European continuum, we can see that this approach performs poorly for distinguishing sub-groups of Indo-European families. For instance, Spanish (Romance language), Norwegian (Germanic language), and Czech

(Slavic language) are placed next to each other. Even worse, Croatian (Slavic language), Portuguese (Romance language), Latin (Latin language), Finnish (Finnic language, does not belong to the Indo-European group), Ancient Greek (Greek language), Slovak (Slavic language), Danish (Germanic language), Basque (Basque language, does not belong to the Indo-European group), Estonian (Finnic language, does not belong to the Indo-European group), Persian (Iranian language), and Hebrew (Semitic language, does not belong to the Indo-European group) are next to each other. Our results are not as satisfying as the results with 20 languages reported by Liu (2010), which rather successfully distinguished sub-groups of Indo-European families. The cause of this difference is hard to identify. Possible explanations can be: 1) the genre differences; 2) the treebank size differences; 3) the approach is good at coarse distinction but it is not decent for more fine-grained task; 4) the inhomogeneous annotation schemes are better suited for difference measures than the rather homogenous UD annotation scheme (Different from common annotations, UD promotes content words, instead of function words, as the heads of dependencies. This strategy may relatively weaken the differences between languages). If we had access to comparable treebanks – or even better to parallel treebanks – which are still a very scarce resource, available for only very few languages at this stage, we decide to leave this question to future research.

Fig. 3 reveals some incoherence of the current state of the UD. It indicates the different places taken by the English (arrows on the left side) and the French (arrows on the right side) treebanks of UD_V2.0. Similar situations arise for Italian, Czech, Russian, etc. We obtain no satisfying results for any multi-treebank language dataset. Note also that the parallel treebanks from the ParTUT team (the Multilingual Turin University Parallel Treebank, Sanguinetti 2013, circled in Fig. 3) coherently have a lower percentage of positive dependency links than their counterparts for English, French, and Italian. It seems that any derived typological classification could remain quite treebank dependent at the current state of UD.

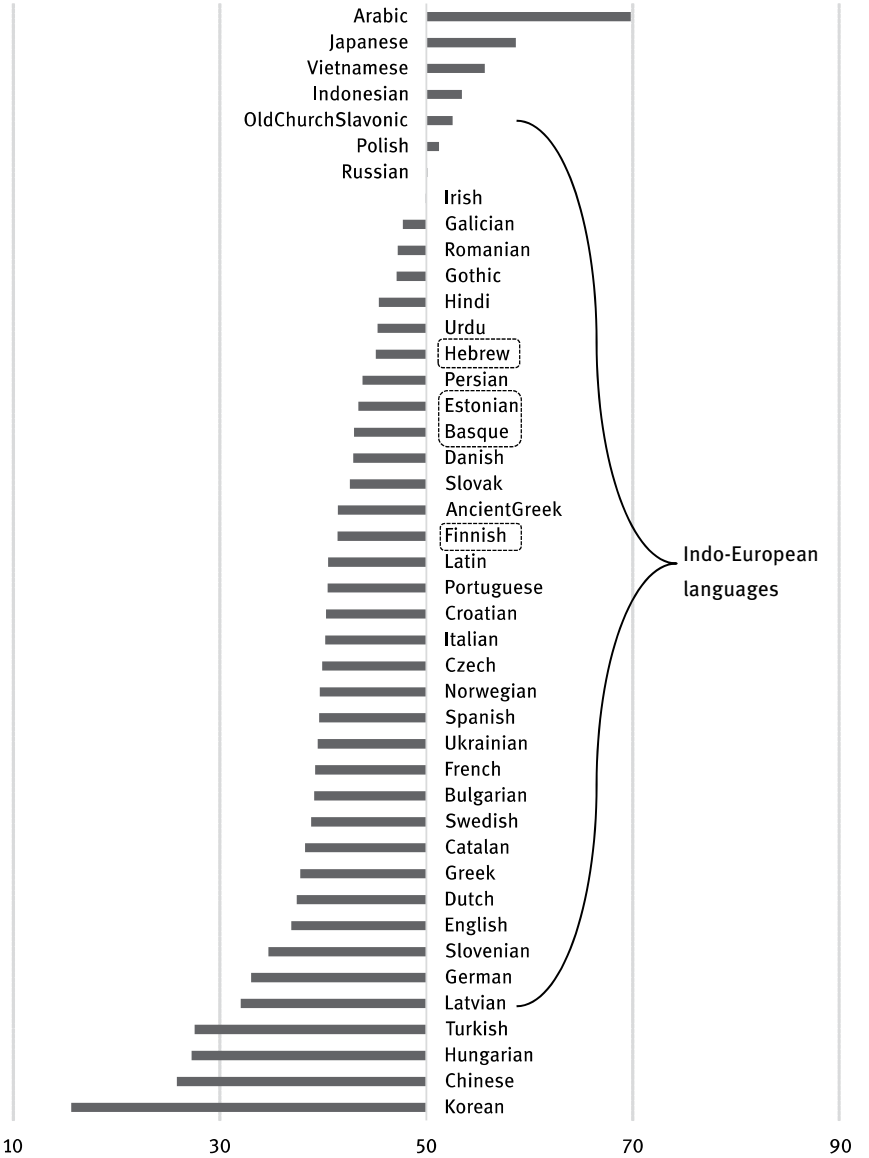


Fig. 2: Languages ordered by % of positive links

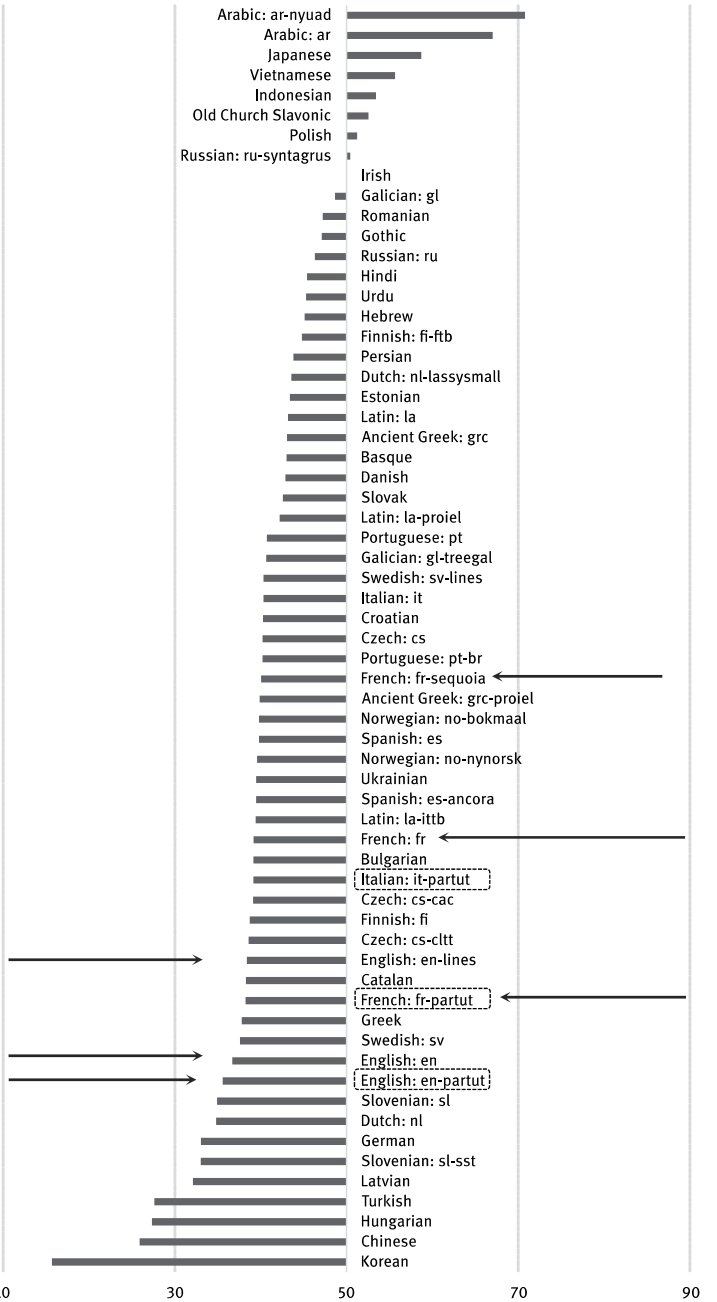


Fig. 3: Treebanks ordered by % of positive links

3.2 DDD of languages and treebanks

The DDD for each language is illustrated in Fig. 4 and corresponding values for treebanks are illustrated in Fig. 5. For detailed values see Appendix 3.

Different from the results of Fig. 2, DDD has a relatively lower accuracy for distinguishing Indo-European languages from other languages. Although the number of languages that manage to make their way into the Indo-European continuum is 4, which is the same with the results of the simple measure of Fig. 2, there are also Indo-European languages, such as Hindi and Urdu, mixed into the other languages. Yet, in general, DDD still does a good job for this coarse distinction task with an accuracy (43 – 7) * 100 / 43 = 83.7%. Besides the broad sense accuracy, some interesting details are worthy of more attention. Observe how Japanese finds its natural position close to Korean, Turkish, and Hungarian (they are all agglutinative languages) in the DDD measure, whereas the direction percentage measure places Japanese right next to Arabic. One should bear in mind that the simple accuracy may not be a decent assessment for deciding which approach is 'better'.

On the other hand, DDD does deal with sub-groups of Indo-European families better. Although the Germanic language group spreads across the spectrum, the Romance languages, however, are very well clustered around an average distance of about 0.8.

Similar to Fig. 3, Fig. 5 also reveals some incoherence of the current state of the UD but with different patterns. It indicates that the different places taken by the English (arrows on the left side) treebanks are more distant than that of the French (arrows on the right side), which is opposite to Fig. 3. Although the absolute values are not as extremely different as the position suggests (en: 0.36, 0.53, 0.77; fr: 0.64, 0.76, 0.79), it confirms that any derived typological classification seems to remain quite Treebank-dependent at the current state of UD. Another detail is also coherent with what we find in Fig. 3. The parallel treebanks from the ParTUT team (circled in Fig. 5) show a similar pattern. They coherently have lower dependency directions than their counterparts for English, French, and Italian. It is tempting to attribute this difference to differences in the guidelines used by different teams in the annotation process. So maybe the difference is rather due to the syntactic structure of "Translationese" (Baroni & Bernardini 2005), that has shorter dependency links for the mostly head-initial languages included in ParTUT. More generally, this shows how these methods also allow for detecting common ground and outliers in the process of treebank development, making them a tool of choice for error-mining a treebank.

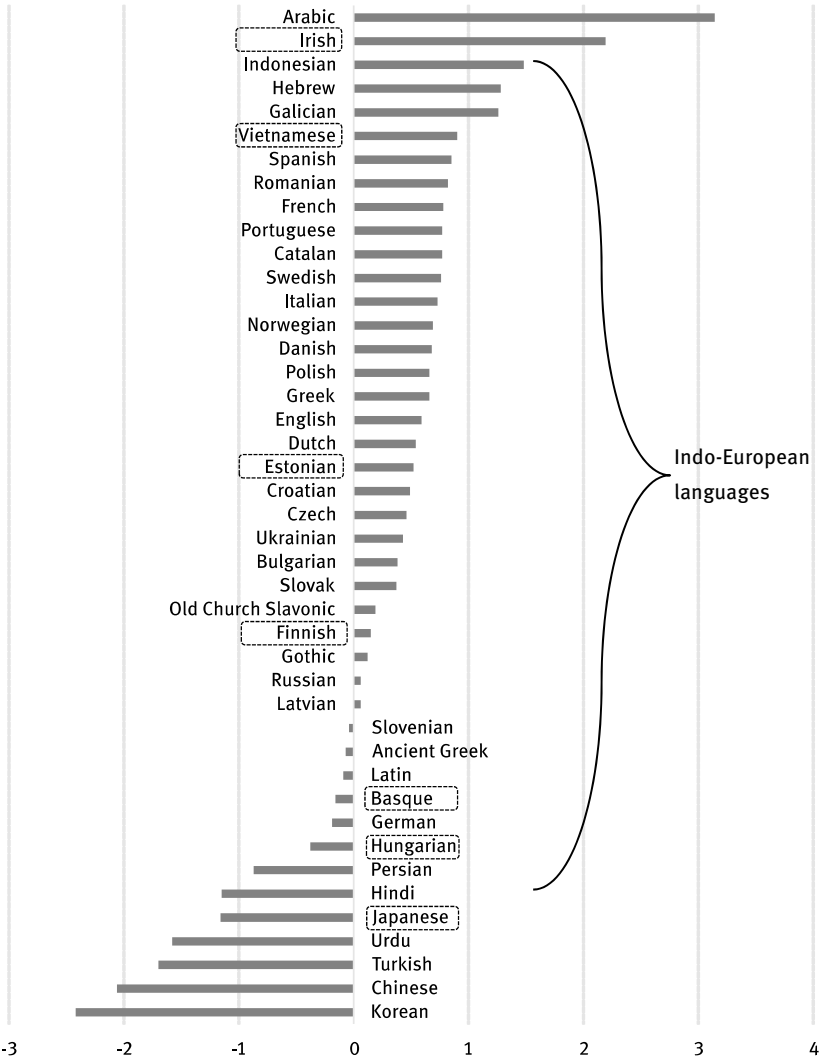


Fig. 4: Languages ordered by directional dependency distance

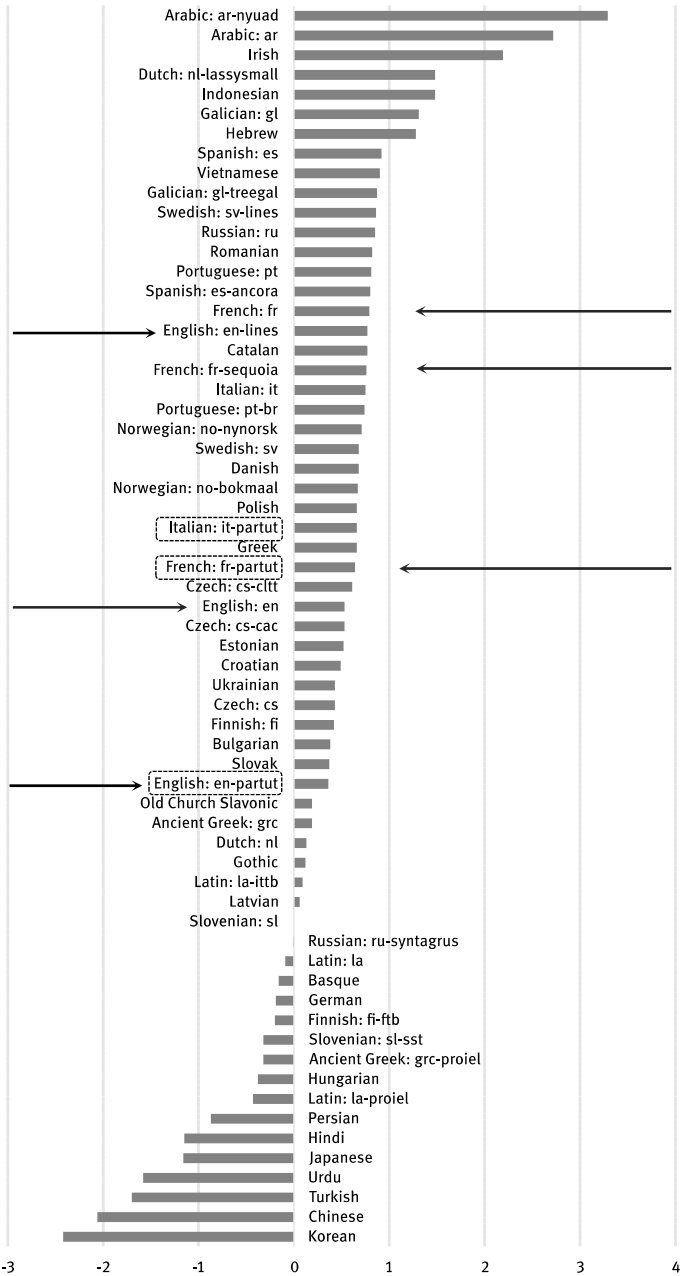


Fig. 5: Treebanks ordered by directional dependency distance

4 Conclusion

Our research shows that Tesnière's idea, which is classifying languages by head-initial and head-final features of all grammatical structures in a sentence, is compatible with empirical analysis.

Liu (2010) completes this idea by measures of the proportions of positive and negative dependency of a language. The measure proposed alongside empirical support of 20 languages is tested in this study with a new dataset, the UD_V2.0 dataset. Our results show that this approach performs very well in a coarse sense but not as satisfying as previous studies (Liu 2010) suggest. Although, the cause of less satisfying results cannot be identified in the current state, we still could cautiously conclude that this approach is proved to be valid.

We also propose and present a new measurement in this study. Results show that this DDD approach performs better in more fine-grained task. It is possible that DDD reflects different typological features in comparison to previous measurement. In general, it is also an effective approach for distinguishing language groups.

Nevertheless, one should not ignore language classification studies that are less 'orthodox', which carry more engineering genres rather than linguistic ones. Many studies seem to obtain more satisfying results with multi parameter data or multiple metrics that provide more details than simple percentages or mean distances (Chen & Gerdes 2017; Liu & Li 2010; Abramov & Mehler 2011; Liu & Xu 2011, 2012; Liu & Cong 2013). However, it remains a job for linguists in the future to improve existing approaches of analysing language data and provide comprehensive explanations for all these phenomena.

Acknowledgement: This work is supported by the National Social Science Foundation of China (Grant No. 18CYY031).

References

Abramov, Olga & Alexander Mehler. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics, 18*(4), 291–336.

Baroni, Marco & Silvia Bernardini. 2005. A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing, 21*(3), 259–274.

Bickel, Balthasar. 2007. Typology in the 21st century: major current developments. *Linguistic Typology, 11*(1), 239–251.

Chen, Xinying & Kim Gerdes. 2017, September. Classifying Languages by Dependency Structure. Typologies of Delexicalized Universal Dependency Treebanks. In *Proceedings of the*

Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy (No. 139, 54–63). Linköping University Electronic Press.

Croft, William. 2002. *Typology and Universals*. Cambridge: Cambridge University Press.

De Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre & Christopher D. Manning. 2014, May. Universal Stanford dependencies: A cross-linguistic typology. In *LREC* (Vol. 14, 4585–92).

Gerdes, Kim, & Sylvain Kahane. 2016. Dependency Annotation Choices: Assessing Theoretical and Practical Issues of Universal Dependencies. In *LAW X (2016) The 10th Linguistic Annotation Workshop*: 131.

Greenberg, Joseph H. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg (Ed.), *Universals of Language* (pp.73–113). Cambridge, MA: MIT Press.

Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie. 2005. *The World Atlas of Language Structures*. Oxford: Oxford University Press.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency treebanks. *Lingua, 120*(6), 1567–1578.

Liu, Haitao & Jin Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin, 58*(10), 1139–1144.

Liu, Haitao & Wenwen Li. 2010. Language clusters based on linguistic complex networks. *Chinese Science Bulletin, 55*(30), 3458–3465.

Liu, Haitao & Chunshan Xu. 2011. Can syntactic networks indicate morphological complexity of a language? *EPL (Europhysics Letters), 93*(2), 28005.

Liu, Haitao & Chunshan Xu. 2012. Quantitative typological analysis of Romance languages. *Poznań Studies in Contemporary Linguistics, 48*(4), 597–625.

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

Nivre, Joakim, & Lars Ahrenberg Željko Agić. 2017. Universal Dependencies 2.0 CoNLL 2017 shared task development and test data. *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics*, Charles University.

Petrov, Slav, Dipanjan Das & Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv*:1104.2086.

Sanguinetti, Manuela, Cristina Bosco, & Leonardo Lesmo. 2013. Dependency and Constituency in Translation Shift Analysis. In *Proceedings of the 2nd Conference on Dependency Linguistics (DepLing 2013)*, 282–291.

Song, Jae Jung. 2001. *Linguistic Typology: Morphology and Syntax*. Harlow: Pearson Education (Longman).

Zeman, Daniel. 2008, May. Reusable Tagset Conversion Using Tagset Drivers. In *LREC*, 28–30.

Appendix

Tab. 1: General information of treebanks used in the study is taken from the Universal Dependencies project page, http://universaldependencies.org.

Language	Numbers of Treebanks	Size in Tokens (k)	Language Family
Ancient Greek	2	211; 202	Indo-European, Greek
Arabic	2	738; 282	Afro-Asiatic, Semitic
Basque	1	121	Basque
Bulgarian	1	156	Indo-European, Slavic
Catalan	1	531	Indo-European, Romance
Chinese	1	123	Sino-Tibetan
Croatian	1	197	Indo-European, Slavic
Czech	3	1,506; 494; 35	Indo-European, Slavic
Danish	1	100	Indo-European, Germanic
Dutch	2	208; 101	Indo-European, Germanic
English	3	254; 82; 49	Indo-European, Germanic
Estonian	1	106	Uralic, Finnic
Finnish	2	202; 159	Uralic, Finnic
French	3	402; 70; 28	Indo-European, Romance
Galician	2	138; 25	Indo-European, Romance
German	1	292	Indo-European, Germanic
Gothic	1	55	Indo-European, Germanic
Greek	1	63	Indo-European, Greek
Hebrew	1	161	Afro-Asiatic, Semitic
Hindi	1	351	Indo-European, Indic
Hungarian	1	42	Uralic, Ugric
Indonesian	1	121	Austronesian
Irish	1	23	Indo-European, Celtic
Italian	2	293; 55	Indo-European, Romance
Japanese	1	186	Japanese
Korean	1	74	Korean
Latin	3	291; 171; 29	Indo-European, Latin
Latvian	1	90	Indo-European, Baltic
Norwegian	2	310; 301	Indo-European, Germanic
Old Church Slavonic	1	57	Indo-European, Slavic

Language	Numbers of Treebanks	Size in Tokens (k)	Language Family
Persian	1	152	Indo-European, Iranian
Polish	1	83	Indo-European, Slavic
Portuguese	2	319; 227	Indo-European, Romance
Romanian	1	21	Indo-European, Romance
Russian	2	99; 1,107	Indo-European, Slavic
Slovak	1	106	Indo-European, Slavic
Slovenian	2	140; 29	Indo-European, Slavic
Spanish	2	549; 431	Indo-European, Romance
Swedish	2	96; 79	Indo-European, Germanic
Turkish	1	58	Turkic, Southwestern
Ukrainian	1	100	Indo-European, Slavic
Urdu	1	138	Indo-European, Indic
Vietnamese	1	43	Austro-Asiatic

Tab. 2: Percentage of positive links for languages and treebanks

Language	% of positive links	Language	% of positive links
Ancient Greek (all)	41.41	Indonesian	53.44
grc	43.04	Irish	49.93
grc-proiel	39.88	Italian (all)	40.17
Arabic (all)	69.81	it	40.33
ar	67.01	it-partut	39.13
ar-nyuad	70.79	Japanese	58.68
Basque	43.00	Korean	15.67
Bulgarian	39.14	Latin (all)	40.47
Catalan	38.26	la	43.16
Chinese	25.89	la-ittb	39.39
Croatian	40.27	la-proiel	42.21
Czech (all)	39.90	Latvian	32.10
cs	40.21	Norwegian (all)	39.67
cs-cac	39.11	no-bokmaal	39.79
cs-cltt	38.59	no-nynorsk	39.55
Danish	42.89	Old Church Slavonic	52.55
Dutch (all)	37.44	Persian	43.81

Language	% of positive links	Language	% of positive links
nl	34.81	Polish	51.23
nl-lassysmall	43.55	Portuguese (all)	40.41
English (all)	36.91	pt	40.70
en	36.70	pt-br	40.20
en-lines	38.38	Romanian	47.22
en-partut	35.58	Russian (all)	50.08
Estonian	43.39	ru	46.31
Finnish (all)	41.37	ru-syntagrus	50.42
fi	38.72	Slovak	42.58
fi-ftb	44.78	Slovenian (all)	34.68
French (all)	39.23	sl	34.92
fr	39.16	sl-sst	33.00
fr-partut	38.24	Spanish (all)	39.60
fr-sequoia	40.05	es	39.78
Galician (all)	47.72	es-ancora	39.44
gl	48.64	Swedish (all)	38.85
gl-treegal	40.64	sv	37.58
German	33.04	sv-lines	40.33
Gothic	47.11	Turkish	27.61
Greek	37.79	Ukrainian	39.45
Hebrew	45.09	Urdu	45.26
Hindi	45.39	Vietnamese	55.65
Hungarian	27.32		

Tab. 3: DDD for languages and treebanks

Language	DDD	Language	DDD
Ancient Greek (all)	−0.07	Indonesian	1.48
grc	0.19	Irish	2.19
grc-proiel	−0.32	Italian (all)	0.73
Arabic (all)	3.14	it	0.75
ar	2.72	it-partut	0.66
ar-nyuad	3.29	Japanese	−1.16
Basque	−0.16	Korean	−2.42
Bulgarian	0.38	Latin (all)	−0.09

Language	DDD	Language	DDD
Catalan	0.77	la	−0.09
Chinese	−2.06	la-ittb	0.09
Croatian	0.49	la-proiel	−0.43
Czech (all)	0.46	Latvian	0.06
cs	0.43	Norwegian (all)	0.69
cs-cac	0.53	no-bokmaal	0.67
cs-cltt	0.61	no-nynorsk	0.71
Danish	0.68	Old Church Slavonic	0.19
Dutch (all)	0.54	Persian	−0.87
nl	0.13	Polish	0.66
nl-lassysmall	1.48	Portuguese (all)	0.77
English (all)	0.59	pt	0.81
en	0.53	pt-br	0.74
en-lines	0.77	Romanian	0.82
en-partut	0.36	Russian (all)	0.06
Estonian	0.52	ru	0.85
Finnish (all)	0.15	ru-syntagrus	−0.01
fi	0.42	Slovak	0.37
fi-ftb	−0.20	Slovenian (all)	−0.04
French (all)	0.78	sl	0.00
fr	0.79	sl-sst	−0.32
fr-partut	0.64	Spanish (all)	0.85
fr-sequoia	0.76	es	0.92
Galician (all)	1.26	es-ancora	0.80
gl	1.31	Swedish (all)	0.76
gl-treegal	0.87	sv	0.68
German	−0.19	sv-lines	0.86
Gothic	0.12	Turkish	−1.70
Greek	0.66	Ukrainian	0.43
Hebrew	1.28	Urdu	−1.58
Hindi	−1.15	Vietnamese	0.90
Hungarian	−0.38		

Yaqin Wang*, Jianwei Yan

A Quantitative Analysis on a Literary Genre *Essay*'s Syntactic Features

Abstract: Depicted as a puzzling literary genre, essay, in relation to novel, is always given enough attention in literary studies. The present study intends to investigate characteristics of essay through a quantitative analysis of its dependency relations. A detailed investigation was carried out on the relationship, on the one hand, between essay and different types of fictions and, on the other hand, among contemporaneous essayists. Corpora used in this study were composed of four literary genres extracted from LOB Corpus as well as an essay genre constructed by essays of four British essayists. Results show that 1) regarding probability distributions of dependency distances, essay shares similarities with fictions; 2) dependency direction is a useful measurement in language typology rather than in genre judgement; 3) the distribution of dependency type *nsubj* may be a useful metric for distinguishing essay from fictions, whereas, essayists seem to share commonalities within this genre.

Keywords: essay; dependency distance; probability distribution; quantitative linguistics

1 Introduction

Depicted as an eccentric (*excentrique* in Bensmaïa 1986: 528) and puzzling literary genre (Chadbourne 1983), essay constantly mesmerizes literary scholars and indeterminacy becomes one of its common traits (De Obaldia 1995; Ferré 2015). One of many reasons for its uniqueness lies in its peculiar form, fragmented, brief and scattered, with "an artful disorder in its composition" (Chadbourne 1983: 149). Essay, in relation to other conventionalized literary genres, especially novel (fiction), is always given enough attention by the academics (Haas 1966; Chadbourne 1983; Klaus & Stuckey-French 2012). The debate on the relationship between essay and fiction remains unsettled. Some scholars define it as a "catch-all term for non-fictional works of limited length" (Scholes & Klaus

Yaqin Wang, Zhejiang University, Hangzhou, P.R.China, wyq322@126.com
Jianwei Yan, Zhejiang University, Hangzhou, P.R.China

https://doi.org/10.1515/9783110573565-015

1969: 46; Harris 1996). Some regard essay as belonging to imaginative literature (Goddard 1951; Chadbourne 1983). German scholars regard essay as the neutral territory between pure science and pure literature (Chadbourne 1983). So, is essay similar to fiction or as scholars have said, it is simply a non-fiction item? This kind of question has rarely been addressed by empirical corpus-based research according to our knowledge. Hence from the perspective of syntactic features (fragmentary and scattered or not), it may (or not) display divergence from other imaginative genres, say, fictions[1]. Furthermore, featured by the particular relation between the author and his/her own readers, different authors "agree to disagree" on what essays are talking about (Tankard 2015). This implies the exclusive influence, if any, a certain author may exert on essay's form and content. How far the effect goes or even, is there in fact any of this effect? Based on these assumptions, this study intends to probe a little deeper into this literary genre from the aspect of syntactic features.

Dependency grammar describes the syntactic relationship between the words concerned. Several important variables are proposed in a syntactic structure based on dependency grammar (Liu 2009). Dependency distance (DD) is the linear distance between the governor and the dependent. The greater the dependency distance is, the more difficult the syntactic analysis of a sentence is (Gibson 1998; Liu 2009). It is used as a useful metric for investigating hidden language universals (Liu 2008; Futrell et al. 2015; Liu et al. 2017). Dependency direction refers to the word order of dependents and governors. It has been a key measure of interest (Hudson 2003; Liu 2010) in the field of cross-linguistic works. Another feature, viz. dependency type, describes the grammatical relations between the governor and the dependent, thus can well capture the syntactic relationship.

The above three features have been investigated in studies of genre comparison. Liu et al. (2009) noted that there exist some differences of dependency direction between conversational and written genres of Chinese. Wang and Liu (2017) found that genre affects dependency distance and dependency direction significantly, but the effect is small. Within the imaginative and literary genres, Hollingsworth (2012) pointed out that grammatical dependency relations contain stylometric information sufficient for distinguishing authorships. Recently, Yan and Liu (2017) compared dependency distances between two English literary works. These studies confirmed the importance of dependency relations in studying genre variation. However, till now, to the best of our knowledge, few studies have been conducted on syntactic features of essay based on dependen-

1 Lexical features and other linguistic levels are also worthy of further study.

cy grammar from a quantitative linguistic view. The present study, therefore, intends to investigate the characteristics of essay through the analysis of its dependency relation based on a quantitative measure.

Distribution patterns have been applied to model DDs in many languages, be it an exponential distribution (Ferrer-i-Cancho 2004) or a power law, i.e., right truncated zeta, right truncated waring (Liu 2007; Jiang & Liu 2015; Lu & Liu 2016; Wang & Liu 2017). Different parameters for different languages or genres for one probability distribution may indicate the variations across languages or genres (Ouyang & Jiang 2017; Wang & Liu 2017). In terms of these probability distributions, right truncated zeta distribution is an important word-frequency distribution (Baayen 2001) and it has been successfully fitted to dependency distance distribution by several studies (Liu 2007; Jiang & Liu 2015; Wang & Liu 2017; Liu et al. 2017). Wang and Liu (2017) pointed out that the smaller parameter value of the right truncated zeta distribution for the *arts* genre signifies their longer dependency distances, a detailed comparison on these parameters was not carried out though. Therefore, the current research intends to employ this distribution and to compare parameters concerned.

In brief, the study will conduct both the intra- and cross-genre comparisons. We first explore the relationship between essay and fictions, including general fiction, mystery and detective fiction, adventure and western fiction, and humor. Then a finer and more detailed investigation within the genre is carried out by studying contemporaneous essays written by different authors.

Three main research questions are as follows:

1) Does essay demonstrate difference from fictions in terms of dependency distance's distribution? Are those of different essayists also different?
2) Do the distributions of dependency direction vary between essay and different types of fictions and essays written by different authors?
3) Do the distributions of dependency type differ between essay and different types of fictions and among various essayists?

The sections are organized as follows: the following section displays the linguistic materials and quantitative methods employed. Results and discussion related to features of essay's dependency relations are then given in Section 3, after which a brief conclusion is drawn.

2 Method and Materials

Dependency grammar, which focuses on the linguistic relationship between two units within a sentence, can be used in syntactic analysis. A dependency relation can be summarized as a binary and asymmetrical relation, graphed with a labeled arc on top of the dependent and the governor (Tesnière 1959; Hudson 1990; Liu 2009). Here is an example of dependency analysis of the sentence *The boy has an apple* in Fig. 1.

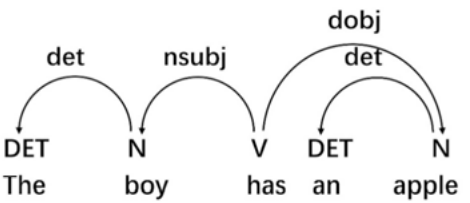


Fig. 1: Dependency analysis result of *The boy has an apple*

As Fig. 1 shows, the arcs labeled with dependency types direct from governors to dependents, which make use of Penn Treebank part-of-speech tags and phrasal labels (De Marneffe & Manning 2008). In the pair (the, boy), for instance, *the* is called the dependent and *boy* is the governor. The labeled arc, defined as *det*, is the type of this dependency relation. The dependency relation, thus, is governor-final.

For computing dependency distance, Jiang and Liu (2015) proposed several methods. Formally, let $W_1...W_i...W_n$ be a word string. For any dependency relation between the words W_x and W_y ($x \geq 1$, $y \leq$ n), if W_x is a governor and W_y is its dependent, then the dependency distance (DD) between them is defined as the difference $x–y$; by this measure, the dependency distance of adjacent words is 1. When x is greater than y, the dependency distance is a positive number, which means the governor follows the dependent; when x is smaller than y, it is a negative number and the governor precedes the dependent. Correspondingly, they are defined as positive dependency and negative dependency respectively in terms of dependency direction.

The mean dependency distance (MDD) of an annotated syntactic corpus can be defined as:

$$\text{MDD (the sample)} = \frac{1}{n-s}\sum_{i=1}^{n-s} |\,\text{DD}_i| \tag{1}$$

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

Stanford parser (version 3.7) was employed to output typed-dependency relations among the words in an input sentence. The output of grammatical relations (De Marneffe & Manning 2008) of the sample sentence by Stanford parser is shown in the following.

det (boy-2, The-1)
nsubj (has-3, boy-2)
root (ROOT-0, has-3)
det (apple-5, an-4)
dobj (has-3, apple-5)

Take the first relation *det* (boy-2, The-1) for example. It indicates that the second word *boy* in the sentence has a dependent *the*, which is the first word of the sentence. The dependency type of this dependency is *det*, or determiner. As for *root* (ROOT-0, has-3), this relation signifies that the root of the sentence is the third word *has* in the sentence. The root relation signifies the position of the head of a sentence instead of a dependency relation and therefore is removed during the computation. Regarding the sample sentence above, the DD of *det* (boy-2, The-1) is 2 − 1 = 1; the DD of *nsubj* (has-3, boy-2) is 3 − 2 = 1; *det* (apple-5, an-4) is 5 − 4 = 1; *dobj* (has-3, apple-5) is 3−5 = −2. Hence, the MDD of the sample sentence *The boy has an apple* in Fig. 1 can be obtained as follows: (|1|+|1|+|1|+|−2|)/4=1.25. Although the parser commonly works rather well (De Marneffe et al. 2006; De Marneffe & Manning 2008), it still makes some errors. When a dependency type is labeled as *dep*, this means the relationship between these two words cannot be determined by the software precisely. Hence, after the preliminary annotation, manual checks and modifications were also carried out.

As for the right truncated zeta distribution, here we introduce the formula based on Liu (2007) and Popescu (2009). First, widely used in linguistic studies, the relative rate of change of frequency ($f(x)$) is assumed to be negatively proportional to the relative rate of change of distance (x). i.e.

$$\frac{df(x)}{f(x)} = -\frac{a}{x}dx \tag{2}$$

The differential equation is then solved to obtain

$$f(x) = \frac{K}{x^a} \tag{3}$$

Since the distance is discrete and texts under investigation are finite, the formula above is transformed into a discrete distribution. Then the normalizing constant K is computed, i.e. we set

$$P_x = \frac{K}{x^a}, x = 1, 2, 3, \ldots, R \tag{4}$$

where the normalizing constant can be written as the sum $K^{-1} = \sum_{x=1}^{R} x^{-a}$. Hence, we obtain

$$P_x = \frac{1}{x^a \sum_{x=1}^{R} x^{-a}}, x = 1, 2, 3, \ldots, R \tag{5}$$

and the R is the point of right truncation. Let us define the function

$$\Phi(b, c, a) = \sum_{x=1}^{\infty} \frac{b^x}{(c+x)^a} \tag{6}$$

since we have $b = 1$, $c = 0$, and the greatest distance is R, Hence the zeta distribution can be written as

$$P_x = \frac{1}{x^a \Phi(1,0,a)}, x = 1, 2, 3, \ldots, R \tag{7}$$

In case of truncation on the right-hand side we have

$$\begin{aligned} \sum_{x=1}^{R} \frac{b^x}{(c+x)^a} &= \sum_{x=1}^{\infty} \frac{b^x}{(c+x)^a} - \sum_{x=R+1}^{\infty} \frac{b^x}{(c+x)^a} \\ &= \sum_{x=1}^{\infty} \frac{b^x}{(c+x)^a} - \sum_{x=1}^{\infty} \frac{b^{x+R}}{(c+R+x)^a} \\ &= \Phi(b, c, a) - b^R \Phi(b, c+R, a) \end{aligned} \tag{8}$$

Hence finally the right truncated zeta distribution can be written as

$$P_x = \frac{1}{x^a[\Phi(1,0,a) - \Phi(1,R,a)]}, x = 1,2,3,\ldots,R \quad (9)$$

Then we use the software Altmann-Fitter (2013) to fit the right truncated zeta distribution to the observed data to test whether the hypothesis is considered as compatible with the data.

For the cross-genre comparison between the English essay and other genres, the Lancaster-Oslo/Bergen Corpus (LOB) was chosen as the data source[2]. The corpus is built to be a representative sample of the texts printed in 1960s. It aims at a general representation of text types, including 15 text categories varying from literature to government reports. In current study, four text categories were chosen as part of the corpus, namely, general fiction (K), mystery and detective fiction (L), adventure and western fiction (N), and humor (R). Ten texts were extracted from K, L, N and nine texts from R (this genre in LOB contains nine texts in total).

Then, we chose four famous British essay writers in the early 20th century whose works are approximately contemporaneous with the LOB Corpus. They are Bertrand Russell (BR), Bernard Shaw (BS), Edward Verrall Lucas (EV) and Max Beerbohm (MB). Ten texts of their essays were collected from Project Gutenberg[3], each composed of around 2,000 tokens. After that, we randomly extracted 10 texts from these four corpora to compose a new corpus, the essay (G), consisting of 20,000 words in total, which is used as a representative category of British essays. Hereafter genre names are represented by their coded categories in LOB and essay names by the authors' name initials. Hence, we have five genre corpora, G, K, L, N and R, and four essay corpora, BR, BS, EV and MB.

During the composition of the corpus, descriptive information in the plain texts without a full stop was deleted, such as authors, headlines of texts, lists etc. The final organization of linguistic materials is shown in Tab. 1, with 79 texts, around 158,000 tokens in total[4].

2 For more information, please refer to http://clu.uni.no/icame/manuals/LOB/INDEX.HTM

3 http://www.gutenberg.org/

4 The calculation of tokens was based on 40 essays written by four authors and 39 texts extracted from LOB. The genre G was not calculated iteratively since it was randomly chosen from 40 essays.

Tab. 1: The composition of corpora

Category	Coded Name	Text number	Token
Genre	G	10	20,628
	K	10	20,819
	L	10	20,892
	N	10	20,772
	R	9	18,492
Author	MB	10	19,655
	EV	10	18,173
	BR	10	22,025
	BS	10	19,374

3 Results and Discussion

3.1 The probability distribution of dependency distance across genres and authors

The distributions of dependency distance of all texts were fitted by right truncated zeta distribution, and results are shown in Tab. 2. As the table shows, the mean values of R^2 presented show that the distributions of dependency distances of all genres and authors are well captured by the distribution ($R^2 > 0.9$).

Fig. 2 displays the fitting result of DD distribution of each genre and each author[5]. It indicates that there exhibits power-law-like relationship between the frequency and the dependency distance among all genres and authors. In other words, the frequency decreases as the dependency distance increases. This tendency contributes to the long-tail distributions shared by different text categories. Combined with Tab. 2, the results reveal that dependency distance distributions of all genres and authors follow the same regularity, supporting the principle of dependency distance minimization (Liu et al. 2017).

5 We present statistics of one random text from each category due to the limited space.

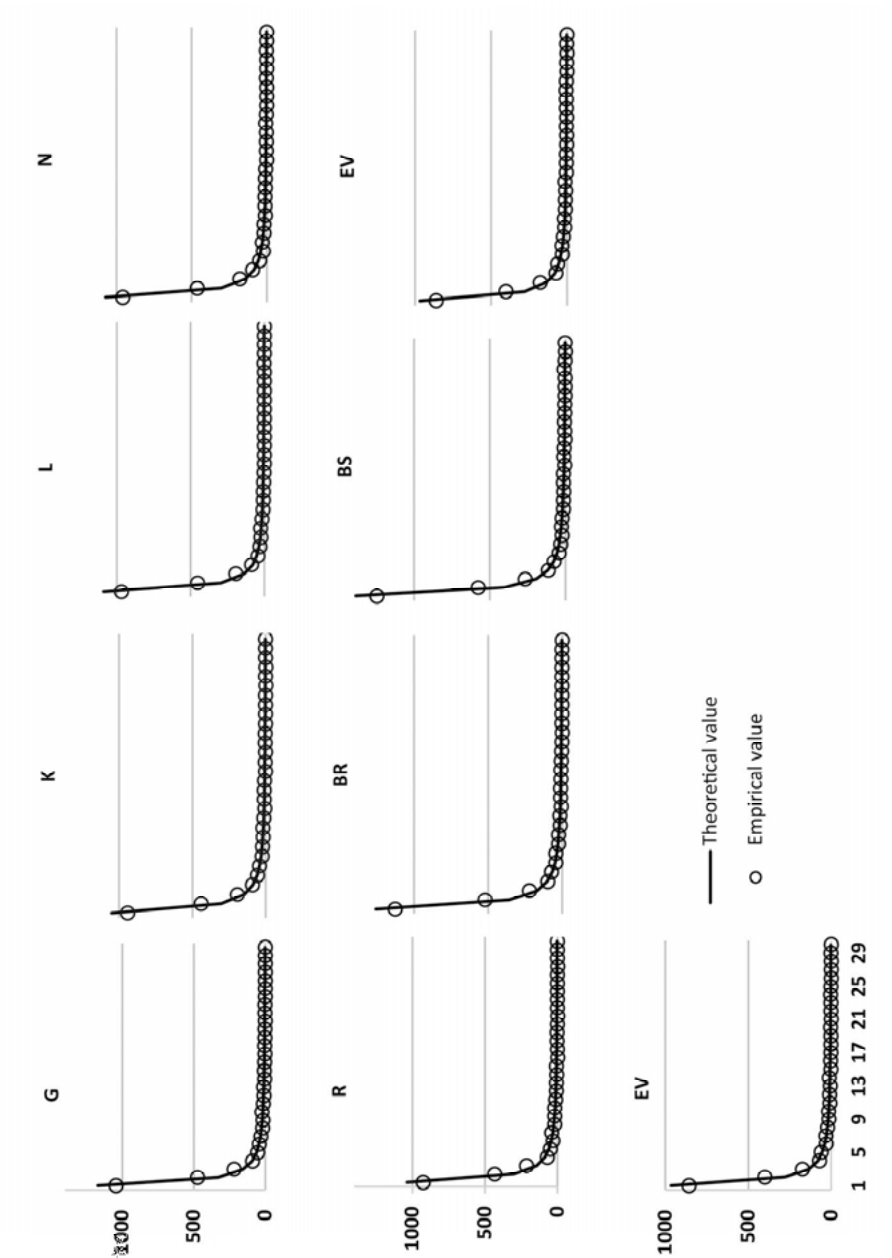


Fig. 2: Fitting right truncated zeta distribution to dependency distances of different genres and authors. The *x*-axis represents dependency distance and the *y*-axis represents DD's corresponding frequency

Tab. 2: Fitting result of right truncated zeta distribution to dependency distance

	Average value of parameter *a*	Average R^2 value
Genre		
G	1.810	0.967
K	1.836	0.960
L	1.838	0.960
N	1.848	0.956
R	1.809	0.963
Author		
BR	1.824	0.970
BS	1.801	0.965
EV	1.777	0.969
MB	1.811	0.968

Tab. 3: Average value of MDD

Genre	
G	3.08
K	2.68
L	2.52
N	2.44
R	2.83
Author	
BR	2.75
BS	3.44
EV	3.31
MB	2.75

Then we focus on the average values of parameter *a* in Tab. 2 and average value of MDD shown in Tab. 3. As Wang and Liu (2017) suggested, smaller value of parameter *a* may signify longer dependencies. Regarding five genres, the average values of parameter *a* for G (1.810) are smaller than those of the other three genres (K, L, N), which has relatively greater MDD value, namely, 3.08 as Tab. 3 shows. This raises a question whether the probability distribution of genre G

significantly differs from that of the other four genres. Hence, a one-way between-subjects analysis of variance (ANOVA) test was conducted on the mean values of parameter a for different genres. The result of the ANOVA test is not significant (F (4, 44) =2.425, p= 0.062). This shows that variation of DD distribution between essay and fictions does not reach a significant level. In other words, the difference between essay and fictions is not obvious and they may share certain similarities. Those essays, "readable but not understandable (Chadbourne 1983; Goddard 1951: 337, cited in Chadbourne 1983)" at the first sight, are not obviously different from imaginative literature. Within this imaginative genre, four essayists' works written in the early 20th century are similar to contemporaneous fictions in terms of dependency distance distribution.

When it comes to four authors, the average value of parameter a for EV (1.777) in Tab. 2 is the smallest among all authors (the other three authors have average values larger than 1.8) with a second greatest MDD value of 3.31 as shown in Tab. 3. A one-way between-subjects ANOVA test was then conducted on mean values of parameter a for different authors. The result shows that there exists significant difference (F (3, 36) = 7.444, p = 0.001). Then the post hoc comparisons using the Tukey's honestly significant difference (HSD) test were followed up. Results in Tab. 4 indicate that the mean parameter for EV (M = 1.777, SD = 0.025) demonstrates a significant difference from those of BR and MB (M = 1.824, SD = 0.027, p = 0.000; M = 1.811, SD = 0.022, p = 0.013, respectively). It shows that only EV is distinct from other authors in terms of DD distribution. Nevertheless, other authors do not display significant differences from each other, signifying DD distributions are close to each other among most authors.

Tab. 4: Results of comparing mean parameters among four authors (p-value)

	BR	BS	EV	MB
BR	—	0.135	**0.000**	0.553
BS	0.135	—	0.105	0.803
EV	**0.000**	0.105	—	**0.013**
MB	0.553	0.803	**0.013**	—

Note: Numbers in bold signify a significant difference (p<0.05)

As mentioned in Introduction, essayists are said to have their own way of controlling the layout of their works; however, they do conform to a certain rule when writing the specific genre, essay. Recalling the aforementioned question

on authorship influence, we may conclude that a certain author does display difference in terms of general DD distributions, however, for most essayists, they somehow share similarities.

3.2 The distribution of dependency direction across genres and authors

In this section, the cross-genre and intra-genre comparisons of dependency directions were conducted in order to figure out the dependency direction distribution patterns of genres as well as those of authors.

Tab. 5: Average percentage of governor-initial and governor-final dependency of different genres

	Proportion of governor-initial dependency	Proportion of governor-final dependency
Genre		
G	0.513	0.487
K	0.489	0.511
L	0.489	0.511
N	0.489	0.511
R	0.503	0.497
Author		
BR	0.508	0.492
BS	0.527	0.473
EV	0.510	0.490
MB	0.490	0.510

The proportions of governor-initial and governor-final dependencies of different genres are presented in Tab. 5. We then plotted the proportion of dependency directions of 10 texts in each genre in Fig. 3 to obtain a better observation.

As Tab. 5 and Fig. 3 show, generally, the variations across genres are kept in a limited range. To be specific, the proportions of negative dependencies in the genre G and R account for a little more than 50%, while the other three genres have similar mean proportions. Besides that, it can be noted that the genre essay (G) has the highest mean proportion of governor-initial dependency (0.513).

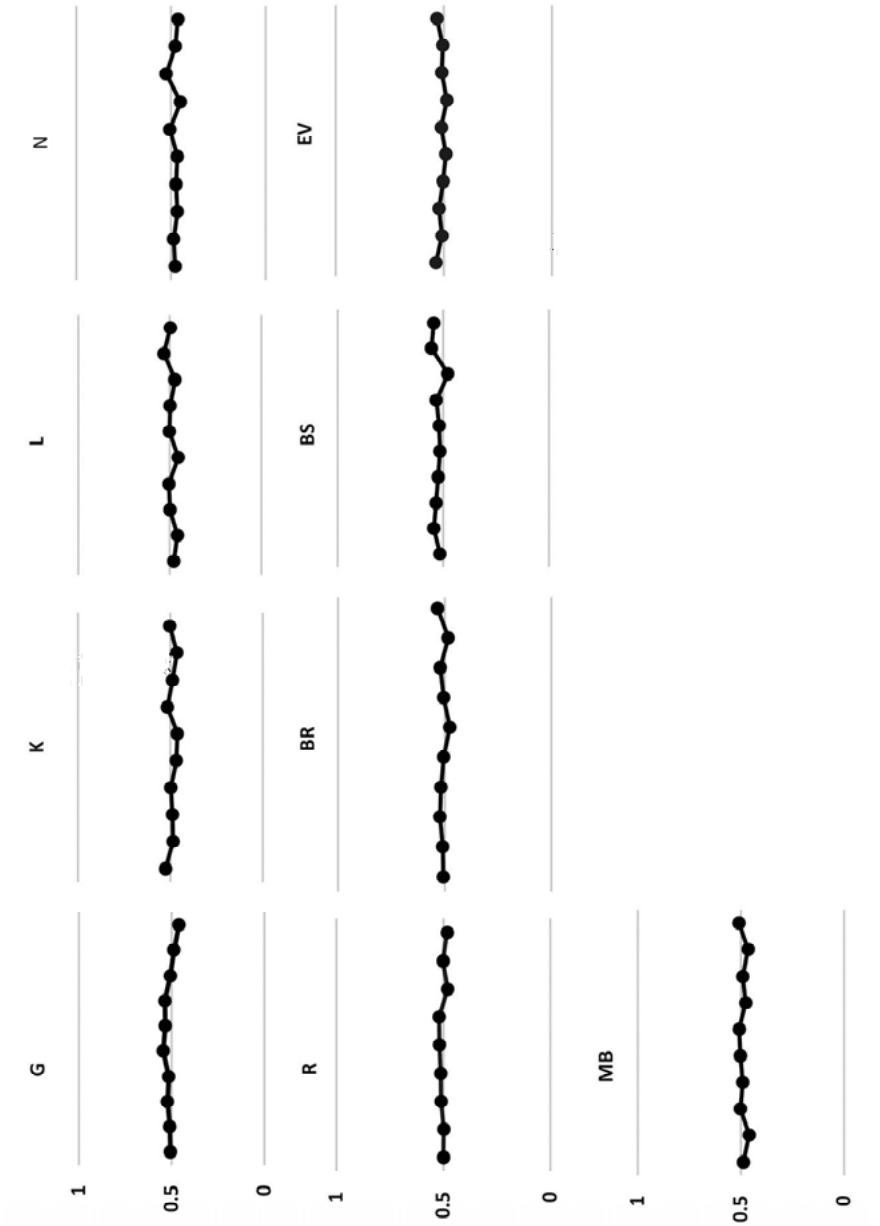


Fig. 3: Proportion of governor-initial dependencies in different genres and authors. The *x*-axis represents texts of a specific genre/author and the *y*-axis represents the proportion of governor-initial dependencies

A Chi-square test was then conducted based on frequencies of negative and positive directions of five genres. The result shows that there is a significant relationship but a very low effect size between genres and dependency direction distributions (χ^2 (4, N = 94497) = 48.390, p < 0.05, Cramér's V = 0.02[6]).

Regarding four essayists, the distributions of dependency directions of all authors also have proportions of negative and positive dependencies that are close to each other. This was also followed by a Chi-square test and an effect size calculation. The results show that there is a significant relationship but again a very low effect size between authors and dependency direction distributions (χ^2 (3, N = 76132) = 53.865, p < 0.05, Cramér's V = 0.03).

Therefore, there are some associations between authors within the genre of essay and the distribution of head-initial dependencies, though the effect of the relationship between the two variables is small. This finding is consistent with previous conclusion on dependency direction that English language tends to have half dependents preceding governors and the other half succeeded by governors (Hudson 2003; Liu 2010) and dependency direction is a useful measurement in language typology rather than in genre judgement (Wang & Liu 2017).

3.3 The probability distribution of dependency type *nsubj* across genres and authors

In the above two sections, we have discussed two major features of dependency relations. In this section, discussion on another feature, i.e., dependency type is carried out. Due to the limited space, a specific dependency type was chosen from all types in the current study. As Biber and Conrad (2009) pointed out, nouns and verbs play an important role in distinguishing genres. Thus we chose a dependency type containing both the nouns and verbs, namely, a nominal subject, abbreviated as *nsubj* (De Marneffe & Manning 2008). The distribution of dependency distance of *nsubj* for each genre and author was fitted by right truncated zeta distribution and relevant results are shown in Tab. 6. The average values of MDD are displayed in Tab. 7.

6 Cramér's V measures the association of two nominal variables (Cramér, 1946). It ranges from 0 to 1. As it can be seen here, Cramér's V is only 0.02, which signifies an extremely low effect.

Tab. 6: Fitting result of right truncated zeta distribution to dependency distance of *nsubj*

	Average value of parameter *a*	Average R^2 value
Genre		
G	1.890	0.980
K	2.095	0.960
L	2.140	0.980
N	2.147	0.958
R	2.042	0.981
Author		
BR	1.897	0.970
BS	1.789	0.974
EV	1.923	0.979
MB	2.003	0.980

Tab. 7: Average value of MDD

Genre	
G	3.14
K	2.15
L	2.00
N	1.95
R	2.22
Author	
BR	2.52
BS	3.63
EV	2.77
MB	2.45

As Tab. 6 shows, the model fitting for all genres as well as authors is quite good ($R^2 > 0.9$). Parameter *a* value displays differences among different texts. For genres, the average value of essay is the lowest compared with other four genres. For authors, except MB, other three authors' values are close to each other. Again, we plotted the empirical values and theoretical values of the model. As Fig. 4 shows, dependency distance distributions of *nsubj* across all texts abide by a power law.

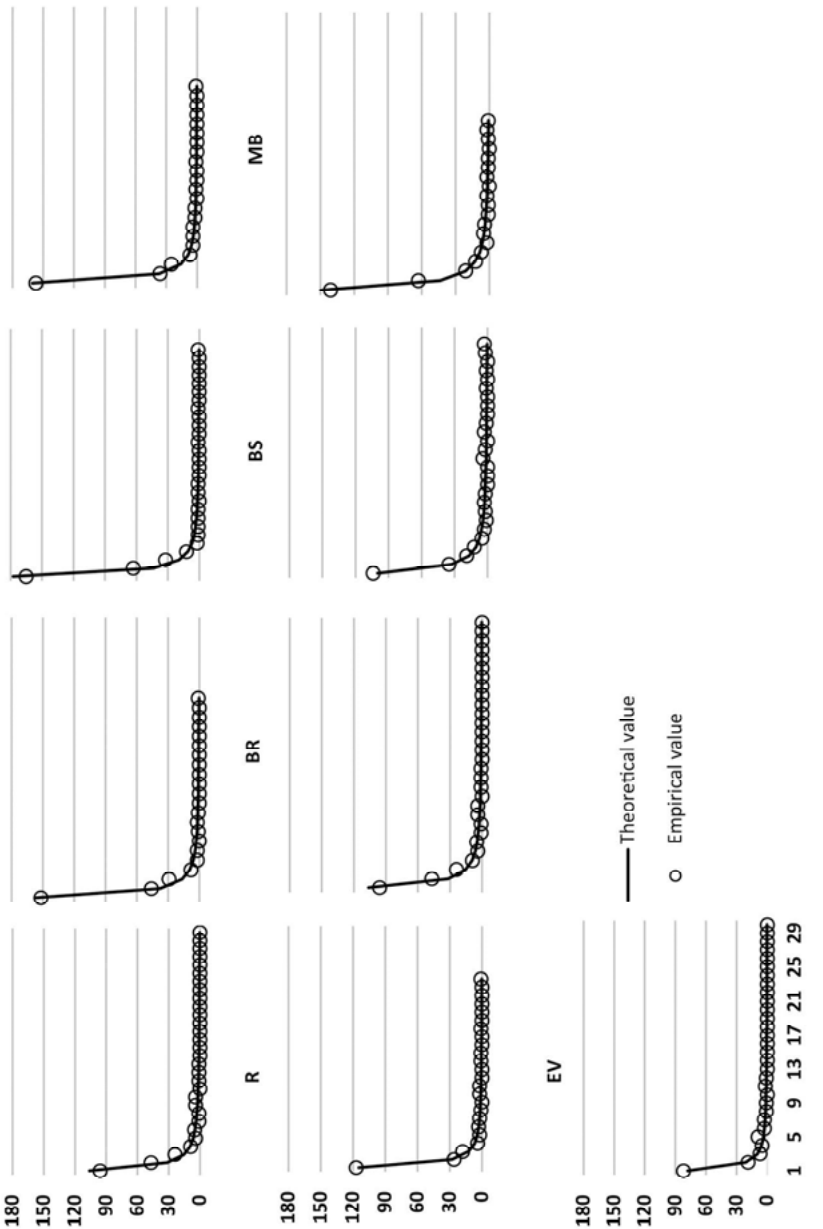


Fig. 4: Fitting right truncated zeta distribution to dependency distances of *nsubj* across different genres and authors. The *x*-axis represents dependency distance and the *y*-axis represents DD's corresponding frequency

We can also, nevertheless, obtain a glimpse of the variations in these distributions. For instance, essay (G) is different from fictions (K, L, and N) for its frequency is smaller than the other three genres when DD equals 1; whereas, it is similar that three of the four essayists, namely, BS, BR, and EV all demonstrate lower frequency.

An ANOVA test was performed on five genres' and four authors' parameters. There is significant difference across genres: F (4, 44) =5.326, p = 0.001. Results of post hoc comparisons using the Tukey HSD test, which is shown in Tab. 8, indicate that a value of essay (M = 1.89, SD = 0.0899) demonstrates significant differences from those of three imaginative genres, namely, K, L and N (M = 2.095, SD = 0.1741, p = 0.04; M = 2.140, SD = 0.0993, p = 0.001, M = 2.147, SD = 0.1750, p = 0.004, respectively).

This means that essay displays significant difference from different types of fictions, including general fiction, mystery fiction, and adventure fiction. The smaller average a value signifies longer distance between the subject and the verb, as shown in Tab. 7 (MDD value is 3.14). This to some extent mirrors previous discussion on essay's famous characteristics, fragmentary and scattered. That is to say, at least from the view of the subject and the verb, it is harder to comprehend than fictions. Combined with what we have discussed in Section 3.1, generally, essay is not significantly different from fictions. In contrast, specifically, certain dependency type can be used for distinguishing different imaginative genres, say, essay from fictions in our case. This empirical finding may be useful for literary studies on essay.

Tab. 8: Results of comparing mean parameters among four authors (p-value).

	G	L	K	N	R
G	—	**0.040**	**0.001**	**0.004**	0.113
L	**0.040**	—	0.738	0.917	0.996
K	**0.001**	0.738	—	0.995	0.527
N	**0.004**	0.917	0.995	—	0.757
R	0.113	0.996	0.527	0.757	—

As for the result of ANOVA test on four authors, their difference is not significant: F (4, 44) =2, p = 0.083. This shows that these four essayists somehow demonstrate similarity in terms of the way of using *nsubj*. Therefore, in a statistical sense, authors' influence does not reach a significant level. Although as a whole, essays are different from another imaginative genre, i.e., fictions in

terms of the usage of *nsubj*, essayists seem to share commonalities within this genre. Divergence across authors may also exists in other types, which may be the scope of the future study.

4 Conclusions

The present study intended to investigate the characteristics of essays through the analysis of its dependency relations based on a quantitative measure. Generally, the probability distributions of dependency distances for both different genres and authors can be well fitted to the right truncated zeta distribution, which means they may follow the similar regularities. Compared with different types of fictions, the genre essay is not significantly different in terms of their probability distribution parameters. In other words, the difference between essay and fictions in terms of dependency distance distribution is not obvious and they may share certain similarities. When it comes to intra-genre comparison, only EV is distinct from other authors in terms of DD distribution. Other authors do not display significant differences from each other, signifying DD distributions are close to each other among most authors. It implies that certain authors do display difference in terms of general DD distributions, however, for most essayists, they somehow share similarities. These findings help clarify the relationship between essay and fictions, or the imaginative literature and may shed light on relevant literary issues.

Both the relationship between the genre and the distribution of dependency direction, between the authorship and the dependency direction are significant, nevertheless the effect sizes of the associations are very small. This again confirms that dependency direction is a useful measurement in language typology (Liu 2010) rather than in genre judgement (Wang & Liu 2017).

As far as dependency distance distribution of *nsubj* is concerned, except humor, essay displays significant difference from general fiction, mystery and detective fiction, and adventure and western fiction. Hence, it is reasonable to conclude that essay is distinctively different from fictions in terms of *nsubj* distribution. In other words, a certain dependency type can be used for distinguishing essay from fictions. This empirical finding may be useful for literary studies on essay. In contrast, four essayists somehow demonstrate similarity in terms of the way of using *nsubj*. Essayists seem to share commonalities within this genre.

The present study explores the relationship between essay and fictions (both similarities and differences can be found), which may be useful for future

literary and stylometric studies. As mentioned before, essay is said to be a meeting ground between pure science and pure literature. It may be a future task to study the relationship between essay and poetry and scientific or informational texts. There may lie something interesting ahead.

References

Altmann-Fitter. 2013. Altmann-Fitter User Guide. The Third Version. Downloadable at. http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf (2014-11-29).

Baayen, Harald R. 2001. *Word Frequency Distributions*. Dordrecht: Kluwer Academic Publisher.

Bensmaïa, R. 1992. Essai. in Demougin, J. (dir). Dictionnaire historique, thématique et technique des littératures: littératures française et étrangères, anciennes et modernes. Paris, Larousse, 1986.

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Chadbourne, Richard M. 1983. A Puzzling Literary Genre: Comparative Views of the Essay. *Comparative Literature Studies, 20*(2), 133–153.

Cramér, Harald. 1946. *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.

De Marneffe, Marie-Catherine & Christopher D. Manning. 2008. Stanford typed dependencies manual (pp. 338–345). Technical report, Stanford University.

De Marneffe, Marie-Catherine, Bill MacCartney & Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 449–454.

De Obaldia, Claire. 1995. *The Essayistic Spirit: Literature, Modern Criticism, and the Essay*. Oxford: Clarendon Press.

Ferré, Vincent. 2015. Aspects de l'essai: références comparatistes, enjeux théoriques.: Journée d'étude de la SFLGC (Société française de littérature générale et comparée), agrégation de Lettres 2016 : " " Inspirations méditerranéennes " : aspects de l'essai au XXe siècle " (Camus, Herbert, Durrell) 13 juin 2015. "Inspirations méditerranéennes" : aspects de l'essai au XXe siècle', Paris, France.

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

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS*, *112*(33), 10336–10341.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Goddard, Harold C. 1951. *The Meaning of Shakespeare*. Chicago: University of Chicago Press.

Haas, Gerhard. 1966. *Studien zur Form des Essays und zu seinen Vorformen im Roman*. Tübingen: Niemeyer.

Harris, Wendell V. 1996. Reflections on the peculiar status of the personal essay. *College English*, *58*(8), 934–953.

Hollingsworth, Charles. 2012. Using dependency-based annotations for authorship identification. In Petr Sojka, Aleš HorákIvan,Ivan Kopeček & Karel Pala (Eds.), *Text, Speech and Dialogue* (pp.314–319). New York: Springer.

Hudson, Richard. 1990. *English Word Grammar*. Oxford: Basil Blackwell.

Hudson, Richard. 2003. The psychological reality of syntactic dependency relations. *MTT 2003, Paris*.

Jiang, Jingyang & 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.

Klaus, Carl H. & Ned Stuckey-French. 2012. *Essayists on the Essay: Montaigne to Our Time*. Iowa City: University of Iowa Press.

Liu, Haitao. 2007. Probability distribution of dependency distance. *Glottometrics*, *15*, 1–12.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, *9*(2), 159–191.

Liu, Haitao. 2009. Probability distribution of dependencies based on Chinese Dependency Treebank. *Journal of Quantitative Linguistics*, *16* (3), 256–273.

Liu, Haitao. 2010. Dependency direction as a means of word-order typology: A method based on dependency Treebanks. *Lingua*, *120*(6), 1567–1578.

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

Liu, Haitao, Yiyi Zhao & Wenwen Li. 2009. Chinese syntactic and typological properties based on dependency syntactic Treebanks. *Poznań Studies in Contemporary Linguistics*, *45*(4), 509–523.

Lu, Qian & Haitao Liu. 2016. Does dependency distance distribute regularly? *Journal of Zhejiang University (Humanities and Social Science),* (4), 63–76.

Ouyang, Jinghui & Jingyang Jiang. 2017. Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*, DOI: 10.1080/09296174.2017.1373991.

Popescu, Ioan-Ioviț. 2009. *Word Frequency Studies*. Berlin: Walter de Gruyter.

Scholes, Robert E. & Carl H. Klaus. 1969. *Elements of the Essay*. Oxford: Oxford University Press.

Tankard, Paul. 2015. Review for *Essayists on the essay: Montaigne to our time*. *Prose Studies*, *37*(1), 92–95.

Tesnière, Lucien. 1959. *Eléments de Syntaxe Structurale*. Paris: Klincksieck.

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

Yan, Jianwei & Siqi Liu. 2017. The distribution of dependency relations in *Great Expectations* and *Jane Eyre*. *Glottometrics*, *37*, 13–33.

Alexander Mehler*, Wahed Hemati, Tolga Uslu, Andy Lücking

A Multidimensional Model of Syntactic Dependency Trees for Authorship Attribution

Abstract: In this chapter we introduce a multidimensional model of syntactic dependency trees. Our ultimate goal is to generate fingerprints of such trees to predict the author of the underlying sentences. The chapter makes a first attempt to create such fingerprints for sentence categorization via the detour of text categorization. We show that at text level, aggregated dependency structures actually provide information about authorship. At the same time, we show that this does not hold for topic detection. We evaluate our model using a quarter of a million sentences collected in two corpora: the first is sampled from literary texts, the second from Wikipedia articles. As a second finding of our approach, we show that quantitative models of dependency structure do not yet allow for detecting syntactic alignment in written communication. We conclude that this is mainly due to effects of lexical alignment on syntactic alignment.

Keywords: dependency structure; authorship attribution; text categorization; syntactic alignment

1 Introduction

In this chapter we introduce a multidimensional model of syntactic dependency trees. Our ultimate goal is to generate fingerprints of such trees to predict the authorship of the underlying sentences. Ideally, one knows the author of a sentence by the kind of tree-like structure spanned by it. However, the shorter the sentence, the smaller the spectrum of candidate structures, the less differentiable such sentences of different authors. Thus, fingerprints of the latter sort are rather a distant goal making it necessary to include other characteristics beyond

Alexander Mehler, Goethe University Frankfurt, Frankfurt, Germany, mehler@em.uni-frankfurt.de
Wahed Hemati, Goethe University Frankfurt, Frankfurt, Germany
Tolga Uslu, Goethe University Frankfurt, Frankfurt, Germany
Andy Lücking, Goethe University Frankfurt, Frankfurt, Germany

https://doi.org/10.1515/9783110573565-016

purely structural ones. The chapter makes, however, a first attempt to create such fingerprints for sentence categorization via the detour of text categorization. To this end, we aggregate the tree-like dependency structures of the sentences of input texts to get aggregated text representations that can be input to authorship attribution. From a methodical point of view our approach integrates quantitative linguistics with machine learning: while the former approach is used to arrive at expressive representations of syntactic structures, the second one is used to perform the desired fingerprinting. In order to test this combined approach, we consider three hypotheses:

Hypothesis 1: Natural language texts are characterized by syntactic alignment so that features of the dependency structure of a sentence correlate with similar features of preceding sentences at a short distance. Though this may relate to the same or different characteristics of sentences, we concentrate on the first alternative: short sentences tend to follow short ones, longer sentences longer ones etc.

Hypothesis 2: Authors are characterized by the way they realize this syntactic alignment: texts by the same author manifest similar patterns of alignment while texts by different authors tend to be different in this respect.

Hypothesis 3: Regardless of whether authors tend towards syntactic alignment or not, they manifest characteristic dependency structures: texts by the same author show similar patterns of dependency structure formation while texts by different authors are different in this respect.

Hypotheses *1* and *2* refer to the notion of syntactic alignment which has been studied mainly in cognitive linguistics (Pickering and Garrod 2004). Hypothesis 3 is a weaker variant which points at the expressiveness of dependency structures in relation to authorship attribution. The chapter addresses all of them. To this end, we start with reviewing related work about syntactic alignment and about quantitative models of dependency structure.

Alignment has mostly been analyzed by example of dialogical communication and is related to priming: priming expressions or constructions (the primes) raise the probability of being repeated in short distance by primed expressions (possibly subject to alterations of the primes). With respect to syntactic dependency structures this means to expect a certain repetition (above chance) of dependency (sub-) trees in proximal contexts. In this sense, a dialog partner manifests, for example, *self-alignment* (Pickering and Garrod 2004) when aligning the syntactic organization of neighboring sentences of her or his speech. We refer to this notion of self-alignment when considering natural language texts of written communication instead of dialogs. Ideally, using measures like autocor-

relation, this kind of repetition effect should be measurable. In particular, however, syntactic priming and alignment proved to be problematic phenomena that are difficult to measure.

Reitter et al. (2006a, b) conceive syntactic priming as the increase of the probability that a syntactic rule is repeated shortly after its use. Syntactic priming has been assessed by means of two kinds of treebanks, the *Switchboard* corpus (Marcus et al. 1993) and the *HCRC Map Task* corpus (Anderson et al. 1991). The authors correlate the probability of repetition of a syntactic rule with the elapsed time between prime and target (carried out in terms of a logistic regression). Without priming, a constant probability is expected, but has not been found, indicating a priming effect. Additionally, the strength of the priming effect differs with respect to the kind of data (it is stronger in *Map Task*) and the direction of processing investigated (it is stronger in the *within speaker*, that is, self-alignment case than in the *between speaker* case).

The verbal alignment model for human-human and human-agent communication of Duplessis et al. (2017) rests on classifying the expressions of dialog partners in terms of *free* (the expression is not part of a larger expression) versus *constrained* (the expression is a subexpression of a larger one). If an expression has been used by both dialog partners and occurred at least once freely, then the expression is said to be *established*. The authors then define *priming* to be the number of repetitions of an expression by the expression initiating dialog partner until the expression is taken up by the other interlocutor. The authors further hypothesize, amongst others, that dialog partners "repeat expressions more often than [expected by] chance" (Duplessis et al. 2017: 75), assessed in terms of the ratio of established expressions and tokens in general. Respective measurements have been carried out by means of negotiation data (Gratch et al. 2016) and compared to randomized variants of the same data. Human-human exchanges have been found to bring about a rich expression lexicon (richer than expected by chance and richer than those of human-agent exchanges) – indicating a sort of alignment.

However, as Howes et al. (2010) point out, many quantitative accounts to structural priming or alignment employ no or only a weak control condition (mainly randomization, see above). Accordingly, they study a particular syntactic construction, namely dative alternation, and compare it to controlled "manipulations" of the same data taken from the *Diachronic Corpus of Present-Day Spoken English*. Control data involve the creation of "fake" dialogs, where either interlocutors (keeping turns fixed) or turns (i.e., sentences; keeping interlocutors fixed) are randomly interleaved. Focusing on a particular, clearly delimited construction, Howes et al. (2010) employ a simple index: a target sentence re-

ceives a score of '1' only if it reuses the dative construction of the most recent prime sentence. The scores are expected to be higher in real dialogs than in "fake" ones. Interestingly, the overall priming scores between the sets of dialogs do not differ significantly. The authors acknowledge that the results, which disagree with many previous findings, may be influenced by the corpora used. However, the findings are consistent with work on lexical alignment and its contribution to structural priming (see e.g. Branigan et al. 2000): "the overall likelihood of a match in syntactic structure across turns appears to be accounted for by the repetition of specific words" (Howes et al. 2010: 2008). Given that the words used reflect what speakers talk about (cf. Pickering and Ferreira 2008), structural similarity may just be an epiphenomenon of the content of dialogs.

Given in particular the evidence for self-alignment found in dialog data (Reitter et al. 2006a, b), there is justification for applying syntactic alignment measures also to monologic text of written communication. In this sense, we hypothesize that self-alignment may also affect the writing process of single and multiple authors. In order to analyze such processes with respect to dependency structure, we need to consider, however, a much wider range of structural features as reference points for measuring syntactic alignment. The reason for this is that instead of counting occurrences of a limited class of expressions, we want to consider a broad range of characteristics of dependency structures at the same time. In the past, such characteristics have been studied in the area of quantitative linguistics. We now briefly review the corresponding literature.

In a series of papers, Ferrer-i-Cancho and colleagues show that the lack of crossing edges in two-dimensional representations of dependency trees is a consequence of short to minimal dependency lengths rather than of an independent principle of syntax (Ferrer-i-Cancho and Gómez-Rodríguez 2016; Ferrer-i-Cancho 2014; Ferrer-i-Cancho and Liu 2014).[1] Ferrer-i-Cancho (2004) conceives the sequencing of syntactically linked words of a sentence as a *minimum linear arrangement problem*. He shows that the Euclidean distances between dependency pairs as found in a Czech and a Romanian dependency treebank are near the theoretical minimum and only grow very slowly (in particular not exponentially) with sentence length. Comparable findings also hold for Chinese (Jiang & Liu 2015).

1 But see, e.g., De Vries (2003), on three-dimensional (constituency) syntax, a representational variant the authors do not consider, which nonetheless gives up the precondition of two-dimensionality necessary for crossing edges.

In a study on 37 languages, Futrell et al. (2015) repeated earlier large-scale studies of Liu (2008). This study corroborated the scope of dependency length minimization in a language-crossing manner. This minimization process concerns the tendency that dependency edges connect adjacent, linearly lined up words. Based on different branching patterns of dependency relations (head-initial vs. head-final and various combinations thereof), the study of Temperley (2008) compared various artificial languages in order to find out the best-minimizing dependency grammar (which is the "one in which the longest dependent phrase of a head branches in the same direction as the head and successive phrases branch on alternate sides" (*op cit.*, p. 280).

Jing and Liu (2015) extend investigations on dependency length (*mean dependency distance* (MDD); see, e.g. Liu (2008) with a hierarchical measure (*mean hierarchical distance* (MHD)) by means of an English and a Czech dependency treebank. The MDD of a sentence is the average distance of the positions of all pairs of dependent and governing vertices within the dependency tree of this sentence. Analogously, the MHD is the average distance of vertices in a dependency tree to its root. Both measures give rise to different distributions in both languages under consideration and are studied in relation to sentence length (with slightly divergent findings for English *vs.* Czech).

A distribution study on the valency and polysemy of 500 high-frequent Chinese verbs, where valence and polysemy information was extracted from a pertinent dictionary, was carried out by Gao et al. (2014). They found that both verb features follow a power-law distribution. In an earlier study, Liu (2011) found, amongst others, that the valency patterns of English verbs follow the positive negative binomial distribution.

A distribution study with a focus on dependency structures was presented by Köhler and Altmann (2000). The authors showed, amongst others, that features of syntactic constructions like length and embedding depth follow a common family of probability distributions. The different kinds of dependencies in Chinese, however, follow, according to Liu (2009) a "modified right-truncated Zipf-Alekseev distribution".

What these approaches have in common is the use of graph-theoretical indices to quantify characteristics of syntactic dependency trees. This is mostly done with a focus on distribution analyses. In this chapter, we adopt the method of quantifying tree-like structures, but instead of distribution analyses we concentrate on classifying textual aggregates based on these quantifications. In this way, we build a bridge between quantitative linguistics on the one hand and machine learning on the other. In the past, related models have already been used to classify documents based on characteristics of their tree-like com-

ponents. Baayen et al. (1996), for example, explore co-occurrence relations in rewrite rules for authorship attribution, while Pustylnikov and Mehler (2007) as well as Mehler et al. (2007) explore characteristics of DOM (*Document Object Model*) trees to classify texts with respect to genre. Further, Abramov and Mehler (2011) consider tree-related indices of syntactic structures in language classification. Next, Mehler et al. (2011) quantify generalized trees to classify Wikipedia-based category graphs. What all these approaches have in common is to quantify tree- or graph-like structures (see, for example, Mehler 2008; Liu & Li 2010; and Mehler 2011) in order to obtain vectors that are finally input to automatic classification or machine learning (Mehler 2008; Macindoe & Richards 2010; Li et al. 2011). The present paper adds such a topology-related approach in terms of quantitative tree analysis and tree classification by drawing on syntactic dependency structures. In this way, we do not only aim at classification (e.g., authorship attribution), but also expect to obtain a broader testbed for measuring syntactic alignment (especially) in (written) communication.

The chapter is organized as follows: Section 2 introduces our multidimensional model of treelike dependency structures. Section 3 describes the data used to instantiate this model. This includes up to a quarter of million sentences taken from (German) texts of two different supergenres (literature and encyclopedia). Section 4 offers a three-part experiment in text categorization, classification analysis and feature analysis to evaluate the model of Section 2. Finally, Section 5 gives a conclusion and a perspective on future work.

2 A Multidimensional model of syntactic dependency structures

In order to test our hypothesis about alignment and the predictability of authorship based on syntactic properties, we experiment with a number of tree invariants[2] and indices that draw on the function-related types of arcs in dependency trees. The former group of characteristics is partly borrowed from preceding work on syntactic structures – especially from Altmann and Lehfeldt (1973), Köhler (1999)[3] and Liu (2008). Our aim is to go beyond these approaches to quantifying dependency trees by simultaneously accounting for several ref-

2 Generally speaking, the value of a graph invariant does not depend on the labeling of its vertices or edges.

3 Note, however, that Köhler (1999) considers constituency instead of dependency structures.

erence points of dependency structure formation. More specifically, we aim to span a multidimensional feature space in order to quantify and compare dependency structures generated from sentences of the same and of different authors. To this end, we quantify the *complexity*-, *dependency*-, *depth*-, *distance*-, *imbalance*-, *length*-, *type*- and *width*-related structure of dependency trees. In order to introduce our multidimensional measurement apparatus, we use the following graph theoretical model of syntactic dependency trees (cf. Kübler et al. 2009):

Definition 1 Let $S = w_0 \dots w_n$, $n \in \mathbb{N}$, be a sentence represented as a sequence of tokens. A Dependency Tree (DT) $T(S) = (V_S, A_S, \iota_S, l_S, r_S)$ representing the syntactic dependency structure of S is a non-empty directed tree rooted in r_S such that $V_S = \{w_0, \dots, w_n\} \neq \emptyset$, $A_S \subset V_S^2$ is the set of arcs $(v, w) \in A_S$ leading from the head v to the dependent w, $\iota_S: V_S \to \{0, n\}$, $\forall w_i \in V_S: \iota_S(w_i) = i$, is a projection function and l_S is an arc labeling function. In order to simplify our formalism, we omit the subscript S and write $T = (V, A, \iota, l, r)$. Further, by $L(V) = \{v \in V | \nexists w \in V: (v, w) \in A\}$ we denote the set of leafs in T. Let $(w_{i_1}, \dots, w_{i_k})$ be the unique directed path starting from w_{i_1} and ending in w_{i_k} in T. Then, the geodesic distance of w_{i_1} and w_{i_k} in T is denoted by $d(w_{i_1}, w_{i_k}) = k - 1$. If there exists no such path between w_{i_1} and w_{i_k}, then $d(w_{i_1}, w_{i_k}) = \infty$. Further, $N_i(v) = \{w \in V | d(v, w) = i\}$ is the set of all vertices at distance i from v in T. Finally, by $degree(v) = |\{w \in V | (v, w) \in A\}|$ we denote the outdegree of v.

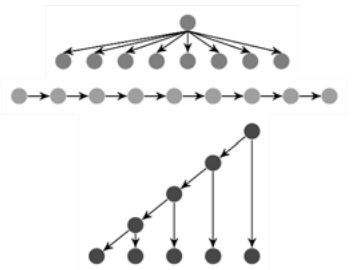


Fig. 1: Top-down: A star graph of depth 1, a line graph of depth 8 and a caterpillar graph each of order 9.

We quantify any dependency tree T along seven meta-dimensions most of which are further quantified along several sub-dimensions:

(1) *Complexity analysis*: by analogy to Köhler (1999), we compute the complexity of a syntactic structure as the ratio of vertices dominated by r (cf. Mehler 2011):

$$comp(T) = \frac{|\{w|(r,w) \in A\}|}{|V|} \in [0,1] \quad (1)$$

comp is related to the valency of verbal roots and the number of (optional) adjuncts by which they are accompanied in S: the higher their number, the more complex the structure of S. Thus, *comp* can be seen as a simple way to characterize authors who tend to manifest complex sentence structures. As a second invariant, we utilize the absolute number of top-level children:

$$\hat{C}(T) = |\{w|(r,w) \in A\}| \in \mathbb{N} \quad (2)$$

Thirdly, we consider the order of T (i.e. sentence length measured by the number of tokens) as a complexity index:

$$order(T) = |V| \in \mathbb{N} \quad (3)$$

The reason for calculating invariants of this simplicity is to perform a parameter study in which we determine whether these alternatives provide the same information as their more complex counterparts (see below).

(2) *Dependency analysis:* we characterize the dependency structure of a tree using the dependency index of Altmann and Lehfeldt (1973):

$$\text{depend}(T) = \frac{2\sum_{i=1}^{depth(T)+1} i|\{v \mid d(r,v) = i-1\}|}{|V|(|V|+1)} \in (0,1] \quad (4)$$

$depth(T)$ denotes the depth of T – see Formula 5. The higher $order(T)$, the more vertices are subordinated in T, the higher the value of depend(T). However, the higher the average distance of vertices to the root, the higher the value of depend(T) in a tree of equal order. In this sense, depend(T) distinguishes star graphs of depth 1 from equal-order line graphs (see Fig. 1). As a second dependency index, we compute the stratum index of Botafogo et. Al (1992): the higher stratum(T), the higher the degree by which vertices are subordinated in T.

(3) *Depth analysis:* to quantify the depth by which T is structured, we consider four invariants:

$$\text{depth}(T) = |\{r, w_{i_j} \dots, w_{i_{j+k}}\}| \in \mathbb{N}\,, w_{i_j} \in L(V) \tag{5}$$

is the length of the longest path $(r, w_{i_j}, \dots, w_{i_{j+k}})$ starting from r. Further, we compute the ratio

$$\hat{T}(T) = \frac{\text{depth}(T)}{order(T)} \in (0,1] \tag{6}$$

to distinguish line graphs, in which the majority of vertices contribute to $\text{depth}(T)$, from higher order trees of the same depth. In order to pay special attention to leafs, we compute the so-called *Leaf Depth Entropy* (LDE) of the set of subsets

$$\mathbb{L} = \{L_i = \{v \in L(V) | d(r,v) = i\} | \, |L_i| > 0\} \tag{7}$$

as

$$LDE(T) = H(\mathbb{L}) = \begin{cases} \dfrac{-\sum_i^{|\mathbb{L}|} p_i \log_2(p_i)}{\log_2|\mathbb{L}|} & : |\mathbb{L}| > 1 \\ 0 & : else \end{cases} \in [0,1], \qquad p_i = \frac{|L_i|}{|L(V)|} \tag{8}$$

$\log_2|\mathbb{L}|$ is the entropy of a same-order set of equiprobable subsets. $LDE(T)$ is maximized by a caterpillar graph of the same order as T, in which no two leafs have the same distance d to the root (other than the one shown in Fig. 1). Thus, LDE can also be seen as a measure of imbalance (see below).

By analogy to the notion of hapax legomena in quantitative lexicology (Tuldava 1998) we additionally compute the ratio $\hat{L}_1$ of vertices w at distance $d(r,w) = 1$:

$$\hat{L}_1 = \frac{L_1}{\max(1, |V| - 1)} \in [0,1] \tag{9}$$

Obviously, this ratio is maximum in a star graph, while it is minimal in a line graph both of order $|V| \to \infty$.

(4) *Distance analysis:* as a measure of the distance structure manifested by dependency trees, we compute the *Mean Dependency Distance* (MDD) of Liu (2008) (see also Jiang & Liu 2015):

$$\mathrm{MDD}(T) = \sum_{(v,w)\in A} |\delta(v,w)| \in \mathbb{R}_0^+ \tag{10}$$

where $\delta(v,w) = \iota(v) - \iota(w)$. While this invariant has already been studied in detail in the literature, we add three measures to capture more information about distances in dependency trees. First and foremost, this concerns the so-called *Dependency Distance Entropy* (DDE) of the frequency distribution of distances $\delta(v,w)$. It computes the relative entropy of the set of subsets

$$\mathbb{D} = \{D_i = \{(v,w) \in A | \delta(v,w) = i\} \subseteq A | |D_i| > 0\} \tag{11}$$

as

$$\mathrm{DDE}(T) = H(\mathbb{D}) = \begin{cases} \frac{-\sum_i^{|\mathbb{D}|} p_i \log_2(p_i)}{\log_2|\mathbb{D}|} & : |\mathbb{D}| > 1 \in [0,1], p_i = \frac{|D_i|}{|A|} \\ 0 & : else \end{cases} \tag{12}$$

A dependency tree which tends to connect neighboring tokens tends to have low distance entropy, while a tree that is interrelating discontinuous tokens at whatever distances tends to have high values of DDE. In this sense, $\mathrm{DDE}(T)$ can be interpreted as a measure of the imbalance of dependency distances. By analogy to our depth analysis, we additionally compute the ratio

$$\widehat{D}_1 = \frac{|D_1|}{\max(1,|A|)} \in [0,1] \tag{13}$$

of arcs between vertices at distance $\delta(v,w) = 1$ and the ratio

$$\widehat{D}_{\{1\}} = \frac{|\{D_i | |D_i| = 1\}|}{\max(1,|A|)} \in [0,1] \tag{14}$$

of distances manifested exactly once. Obviously, a dependency tree of minimal distance entropy, in which tokens w_i are linked to their neighbors w_{i+1}, maximizes $\widehat{D}_1$. On the other hand, by maximizing $\widehat{D}_{\{1\}}$ we also maximize $\mathrm{DDE}(T)$.

(5) *Imbalance analysis:* we utilize the *Total Cophenetic Index* (TCI) of Mir et al. (2013) to assess the imbalance of T. $TCI(T)$ calculates the sum of depths of all lowest common predecessors of all pairs of leafs in T:

$$TCI(T) = \begin{cases} \sum_{v_i, w_j \in L(V), 1 \leq i < j \leq |L(V)|} depth(lcp(v_i, w_j)) & : l' > 3 \\ 0 & : \text{else} \end{cases} \tag{15}$$

$$\in [0, \binom{l'}{3}]$$

where $lcp(v,w)$ is the *lowest common predecessor* of $v, w \in V$, that is, the highest-level predecessor dominating v and w (Mir et al. 2013). Further, l' is the expected number of leafs in a caterpillar graph of order $|T| + k$, $k = |\{v \in V(T) | degree(v) = 1\}|$:

$$l' = |L(V)| + k \tag{16}$$

Using $TCI(T)$, we calculate the variant

$$imbalance(T) = \begin{cases} \frac{TCI(T)}{\binom{l'}{3}} & : l' > 3 \\ 0 & : \text{else} \end{cases} \in [0,1] \tag{17}$$

as a measure of imbalance: the higher its value, the more imbalanced T. So-called caterpillar trees of the same order $|V|$ as T are maximally imbalanced in terms of $imbalance(T)$ (see Fig. 1). Since Formula 17 does not necessarily distinguish between trees of different orders manifesting the same pattern of structuring (e.g., in terms of star graphs), we need to consider additional invariants for measuring imbalance (see below).

(6) *Length analysis:* once more by analogy to Köhler (1999), we compute the length of a syntactic structure as the ratio of the set of leafs $L(V)$ in T:

$$\hat{L}(T) = \frac{|L(V)|}{|V|} \in (0,1] \tag{18}$$

Obviously, the longer the sentence and the higher $\hat{L}(T)$, the shallower its dependency structure. Conversely, the smaller $\hat{L}(T)$, the higher the number of inner nodes, the deeper the dependency tree. In this sense, $\hat{L}(T)$ spans a range between star and line graphs. In addition to this ratio, we compute the number of leafs as a simple invariant (see the comment above):

$$length(T) = L(V) \in \mathbb{N} \tag{19}$$

(7) *Function analysis:* so far, we disregarded arc labeling. In dependency parsing, arcs are typed to code syntactic functions of corresponding subtrees. In this way, one can distinguish, for example, arguments of a predicate by analogy to the valency of the corresponding verb. Thus, by counting arc labels one gets information about the functional structure of a sentence. In order to account for this level of structuring, we employ a range of simple ratios each of which is based on the following schema:

$$\hat{A}_X = \frac{|\{a = (v,w) \in A | l(a) = X\}|}{\max(1, |A|)} \in [0,1] \tag{20}$$

$$X \in \mathbb{X} = \{\text{CJ}, \text{CP}, \text{DA}, \text{HD}, \text{MO}, \text{NK}, \text{OA}, \text{OA2}, \text{OC}, \text{PD}, \text{RC}, \text{SB}\} \tag{21}$$

Since we use the MATE parser (Bohnet et al. 2013), we consider the following subset of values of X: CJ (conjunct), CP (complementizer), DA (dative), HD (head), MO (modifier), NK (negation), OA (accusative object), OA2 (second accusative object), OC (clausal object), PD (predicate), RC (relative clause), SB (subject).[4] The resulting set of 12 characteristics of dependency trees informs us about the extent to which each type contributes to spanning T. As before, we complement these ratios by the following set of frequencies:

$$A_X = |\{a = (v,w) \in A | l(a) = X\}| \in \mathbb{N} \tag{22}$$

(8) *Width analysis:* last but not least, we analyze the width structure of T by computing

$$width(T) = \max\{|N_i(r)| \mid i = 0..\text{depth}(T)\} \in \mathbb{N} \tag{23}$$

in conjunction with the corresponding level as the smallest number i, for which Expression 23 takes its maximum (cf. Mehler 2011):

$$level(T) = arg\ min_i \max\{|N_i(r)| \mid i = 0..\text{depth}(T)\} \in \mathbb{N} \tag{24}$$

Finally, we compute the ratio of vertices belonging to this level as

4 See http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/kanten.html for the complete list of arc labels.

$$\widehat{W}(T) = \frac{|N_{level(T)}(r)|}{|V|} \in (0,1] \tag{25}$$

At first glance, these invariants are hardly informative, since discontinuous vertices of the same level are unlikely to be syntactically related. However, from a cognitive point of view, they allow the assessment of where dependency trees are more broadly structured: near their roots or near their leaves, where the former case may indicate a more complex sentence structuring. We complement this assessment by computing the following Hirsch index (Laniado et al. 2011):

$$\begin{aligned} h\text{-index}(T) = \max\{i \in \{0, \dots, depth(T)\} | \forall 0 \leq j \leq i : \left|N_j(r)\right| - 1 \\ \geq j\} \in \mathbb{N} \end{aligned} \tag{26}$$

By reference to an ideal population pyramid, h-index(T) informs us about how T grows into the direction of its leafs: the higher the value of h-index(T), the longer this pyramidal growth. As before, we enrich this assessment by computing the ratio of vertices contributing to this h-index:

$$\hat{h}(T) = \frac{|\cup_{i=0}^{h\text{-index}(T)} N_i(r)|}{|V|} \in (0,1] \tag{27}$$

By means of h-index we can finally introduce the following measure of imbalance (Mehler et al. 2018a):

$$h\text{-balance}(T) = \frac{h\text{-index}(T)}{\text{depth}(T)} \in (0,1] \tag{28}$$

Line graphs of order $n \to \infty$ are of lowest h-balance, while star graphs of order 3 are of highest h-balance (cf. Fig. 1).

Tab. 1 summarizes all features of dependency trees introduced so far. For each dependency tree and, thus, for each input sentence, this list of 46 features collected by the set

$$\mathbb{F} = \{comp, \dots, h\text{-balance}\} \tag{29}$$

allows for generating a 46-dimensional feature vector. These vectors are then processed according to the following procedure:

Tab. 1: The list of models of dependency trees used to model syntactic alignment

No.	Feature	Description	Eq.	Ref.
1	comp	complexity ratio	1	Köhler (1999)
2	$\hat{C}$	absolute complexity	2	Köhler (1999)
3	order	order of tree	3	
4	depend	dependency index	4	Altmann et al. (1973)
5	stratum	stratum of tree		Botafogo et al. (1992)
6	depth	depth of tree	5	
7	$\hat{T}$	ration of vertices on longest path starting from root	6	
8	LDE	leaf distance entropy	8	
9	$\hat{L}_1$	ratio of leafs at distance one to root	9	
10	MDD	mean dependency distance	10	Liu (2008); Jiang et al. (2015)
11	DDE	dependency distance entropy	12	
12	$\hat{D}_1$	ratio of arcs between adjacent tokens	13	
13	$\hat{D}_{\{1\}}$	ratio of arcs manifesting distances occurring once	14	
14	imbalance	imbalance index	17	
15	$\hat{L}$	ratio of leafs	18	
16	length	number of leafs	19	Köhler (1999)
17	$\hat{A}_X$	ratio of arcs of type *X*	20	
18	A_x	number of arcs of type *X*	22	
19	width	width of tree	23	
20	level	lowest level of maximum width	24	
21	$\hat{W}$	ratio of vertices belonging to the latter level	25	
22	*h*-index	Hirsch index by level	26	Laniado et al. (2011)
23	$\hat{h}$	ratio of vertices contributing to *h*-index	27	
24	*h*-balance	relative *h*-index	28	Mehler et al. (2018a)

1. *Text vectorization:* We start with mapping each time series of sentence-related feature vectors to a vector representing the underlying document x containing s sentences. This is done by means of five aggregation functions: averaging or mean μ, entropy H, relative entropy H_{rel}, distance correlation $dCorr$ (Székely and Rizzo 2009; Kosorok 2009) and autocorrelation R. That is, for each column

$\mathbf{c}_j = (f_{1j}, \ldots, f_{sj})'$ of the sentence-feature matrix of input document x, aggregation function α is used to generate the jth feature value of x. In this way, we arrive at text representation vectors of the sort

$$\alpha(\mathbf{F}) = (\alpha(\mathbf{c}_1), \ldots, \alpha(\mathbf{c}_m)), m = |\mathbf{F}| \tag{30}$$

which are then input to text classification. In the case of averaging and (relative) entropy, this computation is straightforward. In the case of distance correlation, we correlate each column vector with itself. More specifically, the feature value of the ith sentence is correlated with the value of the same attribute of the $(i+1)^{\text{th}}$, the $(i+2)^{\text{th}}$, ..., and the $(i+5)^{\text{th}}$ sentence. In formal terms:

$$dCorr(\mathbf{F}) = dCorr(\mathbf{c}_1) \circ \ldots \circ dCorr(\mathbf{c}_m) \tag{31}$$

where $dCorr(\mathbf{c}_i), i = 1..m$, is a 5-dimensional vector and $\circ$ is the concatenation operator. Thus, using distance correlation as an aggregation function, 46-dimensional sentence feature vectors are mapped onto 230-dimensional text feature vectors. Analogously, in the case of autocorrelation, we refer to the 1st, 2nd and 3rd lag to measure autocorrelation of a dependency tree's attribute value with that of the following three sentences. In this case, we obtain 138-dimensional feature vectors per input document. Auto- and distance correlation are calculated to measure syntactic self-alignment: the stronger the correlation, the higher the alignment in short distance between sentences. In other words, $dCorr$ and R are used to measure the extent to which characteristics of the dependency tree of the i^{th} sentence allow for predicting corresponding values of subsequent sentences.

2. *Classification:* Starting from the text-related feature vectors of the latter step, we perform two kinds of experiments in text classification. Note that our task is *not* to optimize text classification – in fact, authorship attribution is a well-established research area for which solutions exist that even explore syntactic features (cf. Baayen et al. 1996). Rather, our task is to test Hypothesis 1, 2 and 3 of Section 1: we want to know whether authors manifest syntactic alignment and to which degree they are distinguishable by the syntactic dependency structures induced by their sentences. Thus, we utilize text classification just as a scenario for testing our hypothesis. In order to show that our approach is genre-sensitive, we consider authorship attribution and topic detection as two orthogonal classification tasks. Further, we utilize three different classifiers to show that our findings are more or less independent of the underlying machine learning. More specifically, we experiment with *Support*

*Vector Machine*s (SVM) (Cortes and Vapnik 1995), genetic searches on the parameter spaces of these SVMs – henceforth, this procedure is denoted by gSVM[5] –, and recurrent *Neural Network*s (NN) (Hornik et al. 1989). Any classification being computed is evaluated by means of F-scores (i.e., the harmonic mean of precision and recall).

3. *Classification analysis:* unlike traditional approaches to text categorization we additionally compute the permutation $\pi^{\rightarrow}(\mathbb{C})$ of the target classes $\mathbb{C} = \{C_1, \ldots, C_q\}$, $q > 2$, that maximizes our scoring function in the sense that

$$\pi_k^{\rightarrow}(\mathbb{C}) = \begin{cases} \underset{(C_{i_1}, C_{i_2}) \in \mathbb{C}^2 :\, C_{i_1} \prec C_{i_2}}{arg\,max} \; score_{\chi[\cdot]}\left(\left\{C_{i_1}, C_{i_2}\right\}\right) & : k = 2 \\ \pi_{k-1}^{\rightarrow}(\mathbb{C}) \circ \inf\limits_{\prec} \left\{ \underset{C_{i_k} \in \mathbb{C} \setminus [\pi_{k-1}^{\rightarrow}(\mathbb{C})]}{arg\,max} \; score_{\chi[\cdot]}\left(\left[\pi_{k-1}^{\rightarrow}(\mathbb{C})\right] \cup \left\{C_{i_k}\right\}\right) \right\} & : k > 2 \end{cases} \tag{32}$$

$$\pi^{\rightarrow}(\mathbb{C}) = \pi_q^{\rightarrow}(\mathbb{C}) \tag{33}$$

given the classifier χ and its parameter setting $\chi[\cdot]$. $[\pi_{k-1}^{\rightarrow}(\mathbb{C})]$ denotes the set of all classes contributing to the variation $\pi_{k-1}^{\rightarrow}(\mathbb{C})$ and $score$ is a scoring function (in our case F-score). Further, for any $C_{i_1}, C_{i_2} \in \mathbb{C}$: $C_{i_1} \prec C_{i_2} \Leftrightarrow i_1 < i_2$.[6] Computing $\pi^{\rightarrow}$ requires performing $\binom{q}{2} + \sum_{i=1} q - 2$ classifications.[7] The higher the score of $\pi_2^{\rightarrow}(\mathbb{C})$ and the smaller the slope of the curve spanned by the F-scores of $\pi_k^{\rightarrow}(\mathbb{C}), k > 2$, the higher the separability of the classes in $\mathbb{C}$. Hence, $\pi^{\rightarrow}(\mathbb{C})$ may inform about subsets of better separable classes, especially in the case of classification experiments that are difficult to solve.

5 https://github.com/texttechnologylab/GeneticSVM

6 In our case, the mapping of classes C_i onto indices i reflects their lexicographical order.

7 Note that we do not optimize the following function which would require considering all $\sum_{i=1}^{q} i!$ permutations: $\pi_{\max}(\mathbb{C}) = arg\,max_{(C_{i_1}, \ldots, C_{i_q}) \in perm(\mathbb{C})} \sum_{k=1}^{q} score_{\chi[\cdot]}(\{C_{i_1}, \ldots, C_{i_k}\})$

Since Formula 32 optimizes "from left to right", we consider a second alternative that optimizes "from right to left":

$$\pi_k^{\leftarrow}(\mathbb{C}) = \begin{cases} \mathbb{C} & : k = 1 \\ \pi_{k-1}^{\leftarrow}(\mathbb{C}) \backslash \inf_{\prec} \left\{ \underset{C_{i_k} \in \pi_{k-1}^{\leftarrow}(\mathbb{C})}{arg\ max\ score}_{\chi[\cdot]}([\pi_{k-1}^{\leftarrow}(\mathbb{C})] \backslash \{C_{i_k}\}) \right\} & : q > k > 1 \end{cases} \quad (34)$$

$$\pi^{\leftarrow}(\mathbb{C}) = (C_{i_1}, C_{i_2}, \pi_{q-3}^{\leftarrow}(\mathbb{C}) \backslash \pi_{q-2}^{\leftarrow}(\mathbb{C}), \dots, \pi_{q-q+1}^{\leftarrow}(\mathbb{C}) \backslash \pi_{q-q+2}^{\leftarrow}(\mathbb{C})) \text{ where } \pi_{q-1}^{\leftarrow}(\mathbb{C}) = \{C_{i_1}, C_{i_2}\} \text{ and } C_{i_1} \prec C_{i_2} \quad (35)$$

Obviously, $\pi_k^{\rightarrow} \approx \pi_k^{\leftarrow}$, but not necessarily $\pi_k^{\rightarrow} = \pi_k^{\leftarrow}$.

4. *Feature analysis:* finally, we compute networks of all features in Tab. 1. The aim is to investigate redundancies between features caused by habits of the corresponding author to repeat certain sentence structures. To this end, we compute series of feature networks

$$G_\tau = (V, E, dCorr), \tau \in [0,1], V = \mathbb{F}, \forall \{F_i, F_j\} \in E: dCorr(F_i, F_j) > \tau \quad (36)$$

by varying τ in the interval [0,1] in steps of 0,01. In order to measure the redundancy in the resulting networks, we apply four community detection algorithms (*Fast-Greedy* – Clauset et al. 2004; *Infomap* – Lancichinetti and Fortunato 2009; *Multi-level* – Blondel et al. 2008; *Walktrap* – Pons and Latapy 2005) and compute the ratio $C_{\max}(\tau)/|V|$ of the maximum community size $C_{\max}(\tau)$ at level τ and $|V|$: the higher τ and the larger this ratio, the higher the redundancy in the model (at level τ). Intuitively speaking, we seek models manifesting lower redundancy even for smaller $\tau \ll 1$.

This five-step algorithm of (1) generating features of dependency structure, (2) text vectorization, (3) text categorization, (4) classification analysis and (5) feature analysis is tested in Section 4 on the basis of the data of Section 3.

Tab. 2: Corpus 1 sampled from literary data

Author	Documents	Sentences
Arthur Schopenhauer	7	16,902
Franz Kafka	26	28,997
Friedrich Nietzsche	13	48,745
Hugo von Hofmannsthal	14	54,507
Rainer Maria Rilke	5	8,063
Sigmund Freud	5	2,982
Theodor Fontane	5	24,834
Thomas Mann	5	23,716
$\sum$	80	208,746

Tab. 3: Corpus 2 sampled from Wikipedia data

Thematic area	Documents	Sentences
Performing arts (*Darstellende Kunst*)	10	5,198
History (*Geschichte*)	10	2,881
Health (*Gesundheit*)	10	4,224
Humanities (*Humanwissenschaften*)	10	1,440
Literature (*Literatur*)	10	2,856
Philosophy (*Philosophie*)	10	6,657
Psychology (*Psychologie*)	10	3,224
Economy (*Wirtschaft*)	10	10,091
$\sum$	80	36,571

3 Data

Since our approach is text-related, we cannot sample sentences from different texts as done in most approaches to quantifying dependency structures operating on treebanks whose sentences are sampled, for example, from various newspaper articles. Rather, we have to parse texts completely and to consider them exhaustively and separately. Another reason for doing so is that we want

to ensure that our approach is transferable to new corpora that are automatically tagged with respect to dependency structure. To this end, we experiment with two kinds of corpora: texts authored by single authors and texts authored by multiple authors. Tab. 2 summarizes statistics about the literary data explored in our study: we process 208,746 sentences of 80 documents of 8 authors. All documents have been preprocessed using the tool chain of TextImager (Hemati et al. 2016). This includes sentence splitting and lemmatization based on the LAT lemmatizer of Eger et al. (2016), while the MATE parser (Bohnet et al. 2013) is used to tag the dependency structure of each of these sentences. By means of this data we perform syntax-related authorship attribution to test the algorithm of Section 2. With a few exceptions, the documents referred to in Tab. 2 are available via the Project Gutenberg.[8] While these are historical literary texts, most of which were written by single authors in the late 19th and early 20th century, we use a second corpus drawn from the German Wikipedia to obtain a *tertium comparationis*. Tab. 3 summarizes statistics about this data. For each of the 8 different thematic areas considered in our study, we sampled 10 longest Wikipedia articles that were uniquely attributable to exactly one of these areas by means of hyperlinks to the *Dewey Decimal Classification* (DDC) (OCLC 2008) found within the articles. In this way, we obtain a corpus of 36,571 sentences to be processed – obviously, texts of the length of novels are on average longer than encyclopedic articles. For naming the thematic areas in this second experiment, we utilize a subset of the main topic classification of the German Wikipedia[9] extended by categories reflecting the OECD classification of the fields of science and technology (OECD 2007). We call this selection *OECD-oriented Category Selection* (OCS). This second corpus is processed by the same procedure of Section 2 – now under the sign of thematic classification or topic detection, respectively. In a nutshell: by comparing the results obtained for Corpus 1 and 2, respectively, we get insights into the extent to which texts are attributable to authors or to thematic areas by examining the dependency structure of their sentences.

8 http://www.gutenberg.org/

9 Cf. https://de.wikipedia.org/wiki/Kategorie:Sachsystematik

4 Experimentation

4.1 Text Categorization

We start with classifying texts according to the procedure of Section 2 by using the literary data of Section 3. For computing SVMs we experiment with the radial basis function (rbf) kernel and search for optimal parameter configurations. This is done regrding the cost variable c in the interval $[10, \ldots, 1000000]$ and the γ variable in the interval $[0.1, \ldots, 0.000001]$. In the standard experiment, denoted by SVM, all 46 features of Tab. 1 are used for classification. In contrast to this, gSVM seeks for an optimal subset of features to improve the outcome of SVM by means of an evolutionary search. Since this search implements a hill climbing algorithm, gSVM is approximating local optima (i.e., lower bounds of optimal scores). Finally, the NN-based classification scenario uses a feedforward network based on a 300-dimensional hidden layer, a learning rate of 0.001 and is running in 5,000 epochs. In any of these cases, F-score is computed according to the leave-one-out cross validation method (Hastie et al. 2001). SVMs are computed using SVM-Light (Joachims 2002) using the one-against-all training method. For implementing gSVM, we encapsulated SVM-Light into an evolutionary algorithm.[10] The NN is computed by means of Keras (Chollet et al. 2015).

Tab. 4: Experimental results using text-related aggregations of representations of syntactic dependency trees derived from Corpus 1 of literary data

Aggregation function	SVM	gSVM	NN
Mean μ	0.861,4	0.918,9	**0.912,5**
Entropy H	**0.923,8**	**0.942,2**	0.837,5
Relative entropy H_{rel}	0.779,8	0.820,5	0.700,0
Distance correlation dCorr	0.574.7	0.574,7	0.650,0
Autocorrelation R	0.531,6	0.573,5	0.612,5

10 https://github.com/texttechnologylab/GeneticSVM

Tab. 5: Experimental results using text-related aggregations of representations of syntactic dependency trees derived from Corpus 2 of Wikipedia articles

Aggregation function	SVM	gSVM	NN
Mean μ	0.076,09	0.105,26	0.325,0
Entropy H	0.285,20	0.354,83	0.425,0
Relative entropy H_{rel}	0.088,23	0.102,94	0.275,0

The results concerning the literary data are shown in Tab. 4 as a function of the aggregation functions of Section 2. When using entropy for aggregating sentence vectors to get text vectors, we yield the best F-scores with (gSVM: 94%) and without (SVM: 92%) optimization.[11] When using averaging as an aggregation function, the NN performs best. In any event, the alignment-oriented functions $dCorr$ (distance correlation) and R (autocorrelation) perform worst irrespective of the classifier. This indicates that for the lags (1-3) and the distances (1-5) considered here either (1) alignment does not exist, or (2) it is not observable by means of the model of Section 2, or (3) it exists in a way that makes texts of different authors indistinguishable. Since the latter interpretation is contradicted by the very good separation based on entropy, we prefer interpretations (1) and (2). When examining the values of $dCorr$ and R in more detail, we observe that correlation is mostly very low (near zero). This indicates that syntactic alignment (whether existing or not) is either not adequately reflected by our model of dependency structures of Section 2, or by the distances between sentences considered here.[12] Under this regime, we conclude that we do not measure syntactic alignment, even though we have implemented the largest model of dependency structure ever considered in a single study. One reason could be that in written communication – and especially in multi-author texts originating from Wikipedia (s. below) – text revision processes override alignment effects.

11 In the case of relative entropy, 20 features are selected to optimize F-score, in the case of averaging 35 and in the case of entropy 43. In the latter case, $\hat{L}_1$, $length$ and $\hat{A}_X$, $X = \text{MO}$, are deselected.

12 Maybe, alignment exists between sentences at longer distances.

Tab. 6: Errors made by the best performing SVM-based classifier using entropy for aggregation

Document	Classified as
Schopenhauer: Anmerkungen zu Locke und Kant, sowie zu Nachkantischen Philosophen	Fontane
Schopenhauer: Aphorismen	Fontane
Schopenhauer: Nachlass Vorlesungen und Abhandlungen	Nietzsche
Nietzsche: Über Wahrheit und Lüge im außermoralischen Sinne	Kafka
Nietzsche: Der Fall Wagner	Kafka
Hofmannsthal: Die Frau ohne Schatten	Fontane
Kafka: Der Gruftwächter	Hofmannsthal
Kafka: Tagebücher 1910-1923	Schopenhauer
Fontane: Meine Kinderjahre	Schopenhauer
Fontane: Schach von Wuthenow	Mann
Mann: Der Tod in Venedig	Nietzsche
Mann: Gladius Dei	Nietzsche

Tab. 6 lists all errors made by the best performing SVM-based classifier: interestingly, Nietzsche is wrongly attributed as the author of texts written by Thomas Mann (i.a., *The Death in Venice* – a novel including many references to Nietzsche's philosophy) and Arthur Schopenhauer, respectively. That the diaries of Kafka are erroneously attributed to be written by Schopenhauer is possibly due to their aphoristic style.

In our second classification experiment we look at the Wikipedia data of Section 3. The corresponding F-scores are displayed in Tab. 5. We observe a dramatic drop in the scoring. Although the NN-based classifier now always performs best (independent of the aggregation function), each score is far below 50%. Either (1) multiple authorship obscures class membership or (2) the topic cannot be predicted based on syntactic patterns of constituent sentences, or (3) the underlying genre (i.e., knowledge communication) impairs predictability. Note that our Wikipedia corpus contains significantly fewer sentences, a fact that may also affect predictability. Evidently, for deciding between these alternatives we need more experiments far beyond the scope of the present chapter. In any event, this second experiment shows that classifiability by means of the model of Section 2 is not an obvious result and thus illustrates the value of authorship-related classification as performed in the first experiment: in the latter experiment, *dependency structure informs about authorship*.

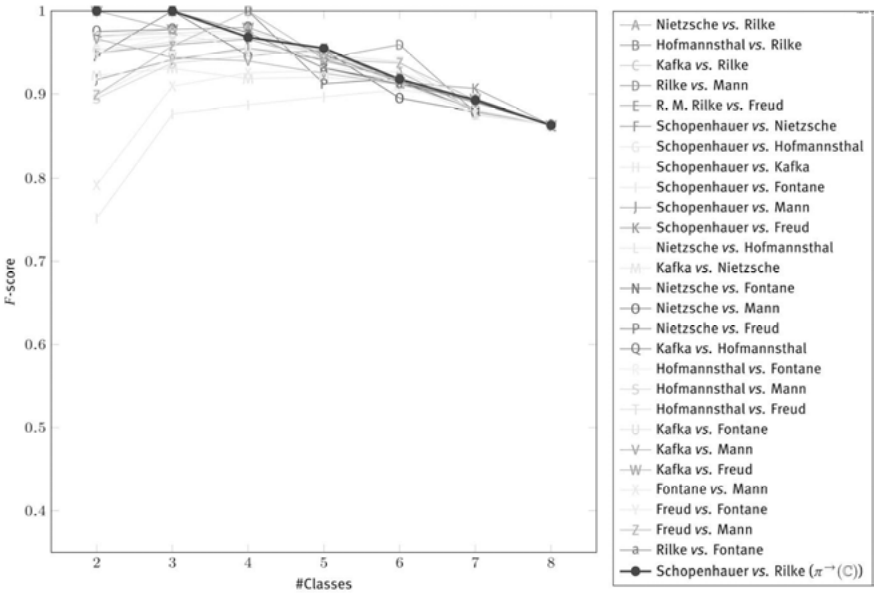


Fig. 2: $\pi^{\rightarrow}(\mathbb{C})$ in conjunction with all other alternatives of seed pairs of authors (literary data). For eight classes (x-axis) the figure displays $\pi^{\rightarrow}_{8}(\mathbb{C})$.

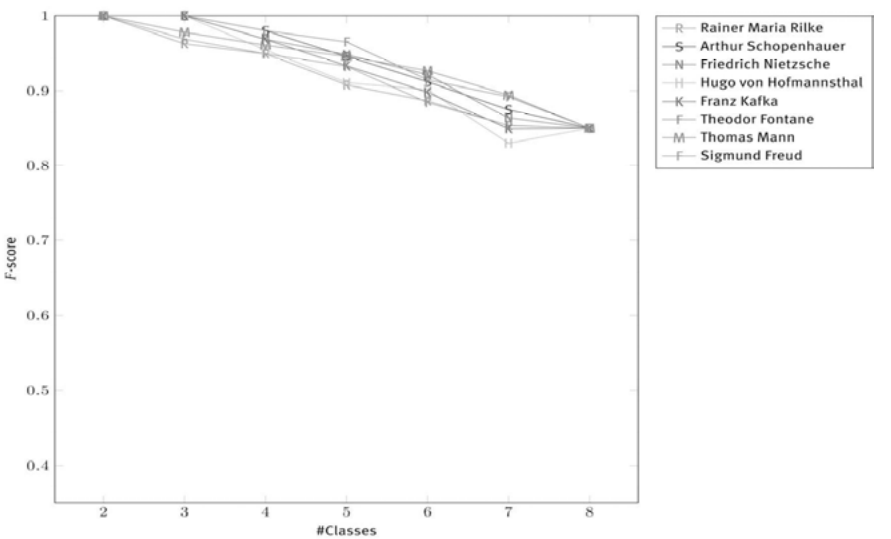


Fig. 3: $\pi^{\leftarrow}(\mathbb{C})$ in conjunction with all alternatives of earliest deselected authors (literary data).

4.2 Classification Analysis

Next, we perform a classification analysis according to Section 2 to get information about subsets of better separable classes. In particular, we want to know the extent to which the best separable pair of authors – denoted by $\pi_2^{\rightarrow}(\mathbb{C})$ – diverge in terms of the syntactic patterns manifested by their sentences. Further, we want to know how to extend $\pi_2^{\rightarrow}(\mathbb{C})$ when trying to optimize F-score according to Formulas 32 and 34, respectively. In this way, our classification analysis determines subgroups of classes whose F-score is higher than the one determined for all classes ($\pi_8^{\rightarrow}(\mathbb{C})$). More specifically, regarding the range of F-scores spanned by all classes (minimum) and $\pi_2^{\rightarrow}(\mathbb{C})$ (maximum), we ask how F-score decreases when approaching $\mathbb{C}$ by gradually increasing the number of target classes starting from $\pi_2^{\rightarrow}(\mathbb{C})$: does it decrease linearly or non-linearly, early or lately, smoothly or abruptly? Obviously, such a classification analysis is more informative than just considering $\mathbb{C}$ as a whole, since it abandons the assumption that all classes in $\mathbb{C}$ are equally separable.

Fig. 2 shows the classification analysis according to Formula 32. It searches for optimal subsets of target classes by gradually increasing their cardinality. It shows that F-score is maximum for three classes. The texts of these authors can be separated without errors by exploring the syntactic patterns of their sentences. There are three alternatives (D, P and Z) showing that this also holds for four classes. However, Scenario D, P and Z start from much lower F-scores than $\pi^{\rightarrow}(\mathbb{C})$. The situation is different if one starts with the pair *Schopenhauer* vs. *Fontane* (I) or *Fontane* vs. *Mann* (X). In this case, F-score increases when the set of target classes is expanded before decreasing to finally reach the score of $\pi_8^{\rightarrow}(\mathbb{C})$. In any event, in most cases, F-score is linearly decreasing when approximating $\pi_8^{\rightarrow}(\mathbb{C})$ thereby starting from a remarkably high level. This demonstrates that F-scores of far above 85% can be achieved for subsets of authors. In other words: while the authors' separability in terms of syntactic patterns is already remarkably high, subsets of them even allow for optimal or near-optimal classifications (F-score $> 95\%$). Note that in each of these classifications we use entropy for aggregating text vectors while training the SVM by means of Weka (Hall et al. 2009) using the same parameter constellation (LibSVM, rbf kernel, $c = 100000$, $\gamma = 1.0 \times 10^{-5}$) for each class and, thus, do not perform any optimization. Further, F-score is computed as the weighted mean of the class-relative F-scores. As a consequence, the F-score of the overall classification is smaller than the one reported in Tab. 4.

The situation is almost the same when examining $\pi^{\leftarrow}(\mathbb{C})$ (see Fig. 3): almost all combinations are linearly increasing (from right to left) thereby rapidly approaching very high F-scores starting from an already high level. Fig. 2 and 3

are contrasted by the classification analysis of Corpus 2 (Wikipedia data) in Fig. 4. Now the F-score of pairs of classes ranges from very low to very high values. For five classes, the average F-score is already close to 50% before it drops below 40%. Thus, while there are pairs of thematic areas that are well separable in terms of the syntactic patterns of their articles (Scenario W (*Philosophy* vs. *Literature*) and Scenario Z (*Philosophy* vs. *Humanities*)), the majority of texts are thematically less separable than in terms of their authorship. Once more our findings indicate that syntactic patterns rather inform about differences of authorship than of the underlying topic.

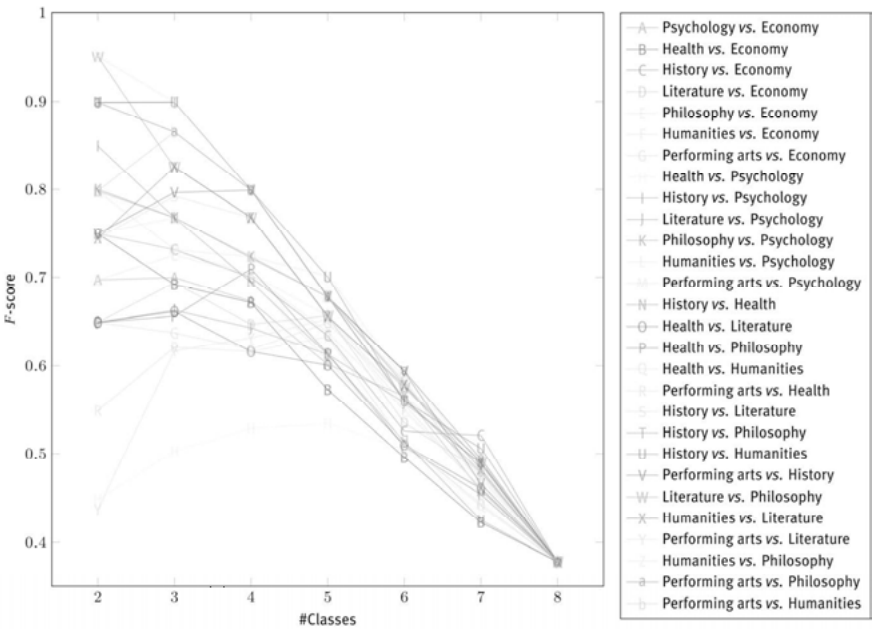


Fig. 4: $\vec{\pi}(\mathbb{C})$ in conjunction with all other alternatives of seed pairs of thematic areas (Wikipedia data). For eight classes (x-axis) the figure displays $\vec{\pi_8}(\mathbb{C})$.

4.3 Feature Analysis

Last but not least, we perform a feature analysis according to the procedure of Section 2: for increasing values of τ in the interval [0,1], we get networks of decreasing density whose community structure is measured as a function of

increasing minimal distance correlation among interlinked features (see Section 2 for the underlying graph model). Fig. 5 and 6 display the results obtained for the works of Kafka and Schopenhauer. The situations illustrated by these authors are very similar: Fast-greedy (a cluster algorithm) and Multi-level (also a cluster algorithm) behave similarly as do Infomap (a flow algorithm) and Walktrap (a cluster algorithm). For $\tau < 0.8$, the latter two methods predict more community formation among redundant features than the former two methods. However, both pairs of community detection algorithms coincide in predicting that this formation rapidly decreases for a distance correlation of at least 80%. In any event, there is remarkably much redundancy in the model even for distance correlations of at least 50% which are known to be high. This indicates a potential for compressing the model of Section 2 in the sense of generating classification results as reported in Sections 4.1 and 4.2 by means of fewer, or, at least, less correlating features.

4.4 Conclusion

Our multidimensional analysis of syntactic dependency structures revealed that authorship correlates with quantitative characteristics of these structures. To show this, we developed a model that allows for predicting the authorship of texts when aggregating sentence-related characteristics of dependency structures: as demonstrated in Section 4.1, these characteristics give access to a kind of fingerprint of authorship. In other words: though we did not measure syntactic alignment as assumed by Hypothesis 1 or 2, we nevertheless observed that dependency structure informs about authorship (Hypothesis 3). To this end, we developed a multidimensional model of dependency trees based on several dozens of characteristics and, thereby, integrated methods of quantitative linguistics and machine learning. We evaluated this model on the basis of a quarter of a million sentences. In any event, our model architecture guarantees the transferability to ever new corpora by using the same parser without having to rely on hand-picked data from treebanks.

Our results about alignment (concerning Hypotheses 1 and 2) are compatible with those of Howes et al. (2010). Given that we have implemented the largest model of dependency structure ever considered in a single study, we subscribe to their suspicion that *syntactic* alignment is highly affected by *lexical* alignment and, ultimately, by semantics (what is talked about). This point can even be strengthened: given that syntax may not be an *autonomous* part of languages, but rather a linguistic tool (Clark 2006) or a sign aspect interfaced with phonology and semantics (Sag 2012), the very notion of syntactic align-

ment, understood as purely syntactic representations influencing each other, is possibly a simplification.[13] Thus, when trying to measure syntactic alignment in written, monologic or dialogical communication, we need to extend our model by including other characteristics beyond purely structural, syntactic ones.

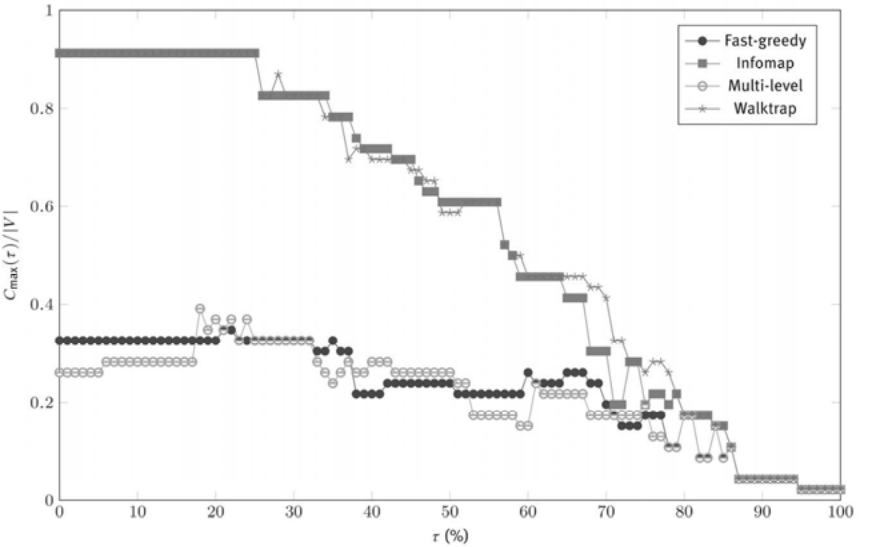


Fig. 5: Community formation among features induced by works of Kafka as a function of minimal distance correlation τ.

13 Note in this respect that Pickering and Garrod (2004) acknowledge syntax as just one level in a network of linguistic features.

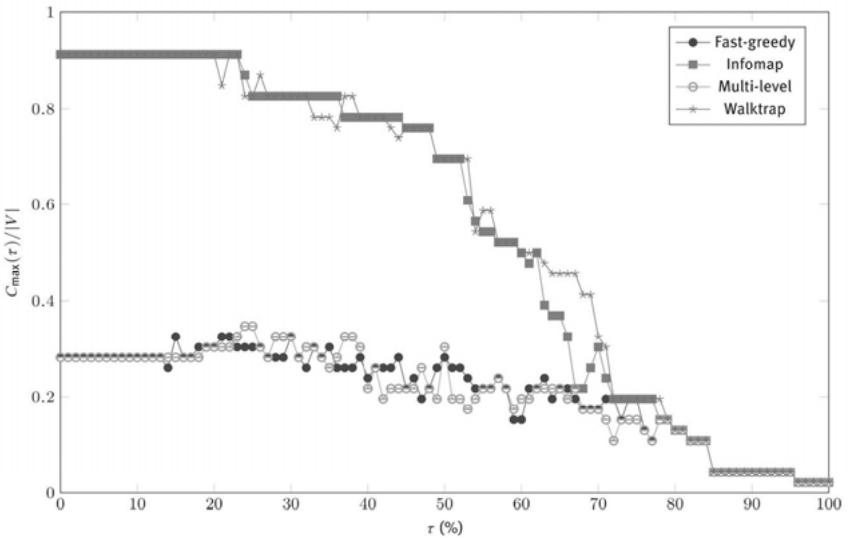


Fig. 6: Community formation among features induced by works of Schopenhauer as a function of minimal distance correlation τ.

Irrespective of the results achieved, our model can be further developed into several directions. First and foremost this concerns the clustering of sentences according to their types and lengths, respectively: up to now, sentences are entered into the aggregation of text vectors on an equal footing – irrespective of their varying lengths (e.g., long vs. short sentences like headers) or types (e.g., assertions vs. questions). The alternative should be to treat sentences of different types and lengths differently so that sentence aggregations of the same type and comparable lengths are compared when classifying different texts. In other words, indices of dependency structure should be made dependent on classifications of sentences.

Secondly, the process by which we selected dimensions for quantifying dependency structures should be systematized. A better approach is to drive forward model generation by means of an axiomatic approach. Candidates for corresponding axioms are classificatory, informational or functional orthogonality. Ideally, the selection process should guarantee that no pair of indices correlates above a certain minimum of allowable distance correlation. In other words, instead of just optimizing F-scores, the compactness and the informativity of the model should be treated as additional dimensions to be optimized. A candidate framework for such a model development is an evolutionary architecture as described by Mehler et al. (2018b).

Thirdly, so far, we did not explore any network information. Thus, the question arises how a valid syntactic network model looks like that accounts for the networking of lexical items (1) of the vocabulary of the input text, (2) of the lexicon of the corresponding author or (3) of the underlying language. That is, beyond the unfolding of dependency structure according to the linear order of sentences, the networking of the vocabulary of a text, author or language can be further reference points of analysis. This may help integrating two still unrelated areas of quantitative linguistics: quantitative models of graphs (e.g., memory models) and of tree-like structures (e.g., hierarchical text or sentence structures).

Further, multidimensional correlations (of groups of dependent and of independent variables) are not yet considered in our model, nor did we look at the full range of linear distances and lags in the linear order of sentences. The reason for extending the model into this direction is to look for more complex access points to syntactic alignment, for which we have not yet found any evidence.

Last but not least, our model should be tested by example of several languages and by means of different parsers to learn its dependency on automatic annotation. As a matter of fact, such parsers make errors. In our case, for example, the MATE parser sometimes creates forests instead of trees.[14] This includes cases in which the parser cannot correctly parse the input sentence or where sentence splitting is incorrect. At present, however, it is unknown what impact such errors have on classifications of the sort carried out here. In any event, using treebanks is not an alternative, as our goal is to quantify complete texts and not just collections of sentences.

References

Abramov, Olga & Alexander Mehler. 2011. Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics, 18*(4), 291–336.

Altmann, Gabriel & Werner Lehfeldt. 1973. *Allgemeine Sprachtypologie*. Fink München.

Anderson, Anne H et al. 1991. The HCRC map task corpus. *Language and Speech, 34*(4), 351–366.

Baayen, Harald, Hans Van Halteren & Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing, 11*(3), 121–131.

14 In all these cases, we explored the largest tree of the forest; the remaining trees mostly consisted of single vertices.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10, P10008.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter & Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, *1*, 415–428.

Botafogo, Rodrigo A, Ehud Rivlin & Ben Shneiderman. 1992. Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems, 10*(2), 142–180.

Branigan, Holly P., Martin J. Pickering & Alexandra A. Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13–B25.

Chollet, F. et al. (2015). Keras. https://github.com/keras-team/keras

Clark, Andy. 2006. Language, embodiment, and the cognitive niche. *Trends in Cognitive Sciences, 10*(8), 370–374.

Clauset, Aaron, Mark EJ Newman & Cristopher Moore. 2004. Finding community structure in very large networks. *Physical Review E, 70*(6), 066111.

Cortes, Corinna & Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning, 20*(3), 273–297.

De Vries, Mark. 2003. Three-dimensional grammar. *Linguistics in the Netherlands, 20*(1), 201–212.

Duplessis, Guillaume Dubuisson , Chloe Clavel & Frederic Landragin. 2017. Automatic Measures to Characterise Verbal Alignment in Human-Agent Interaction. In *Proceedings of the SIGDIAL 2017 Conference*, (pp. 71–81).

Eger, Steffen, Rudiger Gleim & Alexander Mehler. 2016. Lemmatization and Morphological Tagging in German and Latin: A comparison and a survey of the state-of-the-art. In *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016).

Ferrer-i-Cancho, Ramon. 2004. Euclidean distance between syntactically linked words. *Physical Review E, 70*(5), 056135.

Ferrer-i-Cancho, Ramon. 2014. A stronger null hypothesis for crossing dependencies. *EPL (Europhysics Letters), 108*(5), 58003.

Ferrer-i-Cancho, Ramon & Carlos Gómez-Rodríguez. 2016. Crossings as a side effect of dependency lengths. *Complexity, 21*(S2), 320–328.

Ferrer-i-Cancho, Ramon, & Haitao Liu. 2014. The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, 2, 143–155.

Futrell, Richard, Kyle Mahowald & Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *PNAS, 112*(33), 10336–10341.

Gao, Song, Hongxin Zhang & Haitao Liu. 2014. Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics, 21(*1), 1–21.

Gratch Jonathan, David DeVault, Gale Lucas. 2016. The benefits of virtual humans for teaching negotiation. In *Proceedings of the International Conference on Intelligent Virtual Agents in IVA* (pp. 283–294).

Hall, Mark, Eibe Frank, Geoffrey Holmes , Bernhard Pfahringer, Peter Reutemann & Ian H. Witten.2009. The WEKA data mining software: An update. *ACM SIGKDD explorations newsletter, 11*(1), 10–18.

Hastie, Trevor, Robert Tibshirani & Jerome Friedman. 2001. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Berlin/New York: Springer.

Hemati, Wahed, Tolga Uslu & Alexander Mehler. 2016. TextImager: A Distributed UIMA-based System for NLP. In *Proceedings of the COLING 2016 System Demonstrations*.

Hornik, Kurt, Maxwell Stinchcombe & Halbert White. 1989. Multilayer feedforward networks are universal approximators. *Neural Networks, 2*(5), 359–366.

Howes, Christine, Patrcik GT Healey & Matthew Purver. 2010. Tracking lexical and syntactic alignment in conversation. *Proceedings of the Cognitive Science Society*, 32.

Jiang, Jingyang & 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.

Jing, Yingqi & Haitao Liu. 2015. Mean Hierarchical Distance: Augmenting Mean Dependency Distance. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)* (pp. 161–170).

Joachims, Thorsten. 2002. *Learning to Classify Text Using Support Vector Machines*. Boston: Kluwer.

Köhler, Reinhard. 1999. Syntactic structures: Properties and interrelations. *Journal of Quantitative Linguistics, 6*(1), 46–57.

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

Kosorok, Michael R. 2009. On Brownian distance covariance and high dimensional data. *The Annals of Applied Statistics, 3*(4), 1266–1269.

Kübler, Sandra, Ryan McDonald & Joakim Nivre. 2009. *Dependency Parsing: Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Lancichinetti, Andrea & Santo Fortunato. 2009. Community detection algorithms: A comparative analysis. *Physical Review E, 80*(5), 056117.

Laniado, David, Riccardo Tasso, Yana Volkovich & Andreas Kaltenbrunner. 2011. When the Wikipedians Talk: Network and Tree Structure of Wikipedia Discussion Pages. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*.

Li, Geng, Murat Semerci, Bülent Yener & Mohammed J. Zaki. 2011. Graph Classification via Topological and Label Attributes. In *Proceedings of the 9th International Workshop on Mining and Learning with Graphs (MLG)*, San Diego, USA.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.

Liu, Haitao. 2009. Probability distribution of dependencies based on Chinese Dependency Treebank. *Journal of Quantitative Linguistics, 16*(3), 256–273.

Liu, Haitao. 2011. Quantitative properties of English verb valency. *Journal of Quantitative Linguistics, 18*(3), 207–233.

Liu, Haitao & Wenwen Li. 2010. Language clusters based on linguistic complex networks. *Chinese Science Bulletin, 55*(30), 3458–3465.

Liu, Haitao, Chunshan Xu & Junying Liang. 2015. Dependency length minimization: Puzzles and Promises. *Glottometrics*, 33, 35–38.

Macindoe, Owen & Whitman Richards. 2010. Graph Comparison Using Fine Structure Analysis. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing (SOCIALCOM '10)* (pp. 193–200), IEEE Computer Society, Washington, DC.

Marcus, Mitchell P, Mary Ann Marcinkiewicz & Beatrice Santorini. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics, 19(*2), 313–330.

Mehler, Alexander. 2008. Structural similarities of complex networks: A computational model by example of wiki graphs. *Applied Artificial Intelligence, 22*(7&8), 619–683.

Mehler, Alexander. 2011. Social ontologies as generalized nearly acyclic directed graphs: A quantitative graph model of social ontologies by example of Wikipedia. In Matthias Dehmer, Frank Emmert-Streib & Alexander Mehler (Eds.), *Towards an Information Theory of Complex Networks: Statistical Methods and Applications* (pp.259–319). Boston/Basel: Birkhäuser.

Mehler, Alexander, Peter Geibel & Olga Pustylnikov. 2007. Structural classifiers of text types: Towards a novel model of text representation. *Journal for Language Technology and Computational Linguistics, 22*(2), 51–66.

Mehler, Alexander, Olga Pustylnikov & Nils Diewald. 2011. Geography of social ontologies: Testing a variant of the Sapir-Whorf Hypothesis in the context of Wikipedia. *Computer Speech and Language, 25*(3), 716–740.

Mehler, Alexander, Rüdiger Gleim, Andy Lücking,Tolga Uslu & Christian Stegbauer. 2018a. On the self-similarity of Wikipedia talks: A combined discourse-analytical and quantitative approach. *Glottometrics, 40*, 1–45.

Mehler, Alexander, Wahed Hemati, Rüdiger Gleim, & Daniel Baumartz. 2018b. VienNA: Auf dem Weg zu einer Infrastruktur für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprache. In Henning Lobin, Roman Schneider & Andreas Witt (Eds.), *Digitale Infrastrukturen für die germanistische Forschung* (pp.149–176). Berlin: De Gruyter.

Mir, Arnau, Francesc Rosselló & Lucı´ a Rotger. 2013. A new balance index for phylogenetic trees. *Mathematical Biosciences, 241*(1), 125–136.

OCLC (2008). *Dewey Decimal Classification summaries. A Brief Introduction to the Dewey Decimal Classification*. http://www.oclc.org/dewey/resources/summaries/default.htm [accessed February 15, 2009].

OECD (2007). Revised Field of Science and Technology (FOS). www.oecd.org/science/inno/38235147.pdf.

Pickering, Martin J & Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological bulletin, 134*(3), 427–459.

Pickering, Martin J & Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(2), 169–226.

Pons, Pascal & Matthieu Latapy. 2005. Computing communities in large networks using random walks. In pInar Yolum,Tunga Güngör, Fikret Gürgen, Can Özturan (Eds.), Computer and Information Sciences - ISCIS 2005 (pp.284–293). New York: Springer.

Pustylnikov, Olga, & Alexander Mehler. 2007. Structural differentiae of text types: A quantitative model. In Proceedings of the 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKl) (pp. 655–662).

Reitter, David, Frank Keller & Johanna D. Moore. 2006a. Computational Modeling of Structural Priming in Dialogue. In Proceedings of the Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics (HLT/NAACL) (pp. 121–124).

Reitter, David, Johanna D. Moore & Frank Keller. 2006b. Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society (CogSci'06)* (pp. 685–690).

Sag, Ivan A. 2012. Sign-based construction grammar: An informal synopsis. In Hans C. Boas & Ivan A. Sag (Eds.), *Sign-Based Construction Grammar in CSLI Lecture Notes* (pp.69–202). Stanford, CA: CSLI Publications.

Székely, Gábor J& Maria L. Rizzo. 2009. Brownian distance covariance. *The Annals of Applied Statistics, 3*(4), 1236–1265.

Temperley, David. 2008. Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics, 15*(3), 256–282.
Tuldava, Juhan. 1998. *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Trier: Wissenschaftlicher Verlag.

Subject Index

https://doi.org/10.1515/9783110573565-017

Author Index

https://doi.org/10.1515/9783110573565-018

List of Contributors

Beliankou, Andrei
Computerlinguistik und Digital Humanities, University of Trier, FB II / Computerlinguistik, DE-54286 Trier, Germany
E-mail: belianko@uni-trier.de

Čech, Radek
Department of Czech Language, Faculty of Arts, University of Ostrava, Reální 5, Ostrava, 701 03, Czech Republic
E-mail: cechradek@gmail.com

Chai, Gaiying
School of Foreign Languages, Zhejiang Gongshang University, No.18, Xuezheng Str., Xiasha University Town, Hangzhou, 310018, China
E-mail: chaigaiying1@126.com

Chen, Xinying
School of Foreign Studies, Xi'an Jiaotong University, Xianning West Road 28, 710049 Xi'an, Shaanxi, China
E-mail: chenxinying@mail.xjtu.edu.cn

Gerdes, Kim
ILPGA, LPP (CNRS), Sorbonne Nouvelle, 19 rue des Bernardins, F 75005 Paris, France
E-mail: kim@gerdes.fr

Hemati, Wahed
Faculty of Computer Science and Mathematics, Goethe University Frankfurt, Robert-Mayer-Straße 10, D-60325 Frankfurt am Main, Germany
E-mail: hemati@em.uni-frankfurt.de

Hudson, Richard
Department of Linguistics, University College London, Gower Street, London, WC1A 6BT, United Kingdom
E-mail: r.hudson@ucl.ac.uk

Jiang, Jingyang
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China. E-mail: jjy203@163.com

Jin, Huiyuan
Foreign Language Department, Zhengzhou University, No.100 Science Avenue, Zhengzhou, Henan Province, 450001, P. R. China
E-mail: 502209458@qq.com

Köhler, Reinhard
Computerlinguistik und Digital Humanities, University of Trier, FB II / Computerlinguistik, DE-54286 Trier, Germany
E-mail: koehler@uni-trier.de

Koščová, Michaela
Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Mlynská dolina, Bratislava, 84248, Slovakia
E-mail: michaela.koscova@fmph.uniba.sk

Lin, Yanni
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: linyn@zju.edu.cn

Liu, Haitao
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: htliu@163.com

Liu, Jinlu
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail:hardeningwings@163.com

Lopatková, Markéta
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, Praha 1, 118 00, Czech Republic
E-mail: lopatkova@ufal.mff.cuni.cz

Lücking, Andy
Faculty of Computer Science and Mathematics, Goethe University Frankfurt, Robert-Mayer-Straße 10, D-60325 Frankfurt am Main, Germany
E-mail: luecking@em.uni-frankfurt.de

Lu, Qian
College of Language and Literature, Yucai Campus, Guangxi Normal University, No. 15 Yucai Road, Qixing District, Guilin, 541004, China
E-mail: luqian.cn@gmail.com

Mačutek, Ján
Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Mlynská dolina, Bratislava, 84248, Slovakia
E-mail: jmacutek@yahoo.com

Mehler, Alexander
Faculty of Computer Science and Mathematics, Goethe University Frankfurt, Robert-Mayer-Straße 10, D-60325 Frankfurt am Main, Germany
E-mail: mehler@em.uni-frankfurt.de

Milička, Jiří
Institute of the Czech National Corpus, Faculty of Arts, Charles University, nám. Jana Palacha 2, Praha, 1, 116 38, Czech Republic
E-mail: jiri@milicka.cz

Ninio, Anat
Department of Psychology, The Hebrew University, Jerusalem, 91905, Israel
E-mail: anatni@savion.huji.ac.il

Sanada, Haruko
Faculty of Economics, Rissho University, 4-2-16, Osaki, Shinagawaku, Tokyo 141-8602, Japan
E-mail: hsanada@ris.ac.jp

Ouyang, Jinghui
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: oyjh@zju.edu.cn

Uslu, Tolga
Faculty of Computer Science and Mathematics, Goethe University Frankfurt, Robert-Mayer-Straße 10, D-60325 Frankfurt am Main, Germany
E-mail: uslu@em.uni-frankfurt.de

Wang, Hua
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: wanghuazju@163.com

Wang, Yaqin
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: wyq322@126.com

Xu, Chunshan
School of Foreign Studies, Anhui Jianzhu University, 292 Ziyun Road, Hefei, 230601, China
E-mail: adinxu@126.com

Yan, Jianwei
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: yanjianwei@aliyun.com

Yan, Jingqi
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: jqyan@zju.edu.cn

Zhang, Hongxin
Department of Linguistics, School of International Studies, Zhejiang University, No. 866 Yuhangtang Road, Hangzhou, 310058, China
E-mail: mariazhang@yeah.net