Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 21 (2017) 171–193

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Review

Dependency distance: A new perspective on syntactic patterns in natural languages

Haitao Liu [a,b,d], Chunshan Xu [c,*], Junying Liang [b,*]

[a] *Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, 510420, China*
[b] *Department of Linguistics, Zhejiang University, Hangzhou, 310058, China*
[c] *School of Foreign Studies, Anhui Jianzhu University, Hefei 230601, China*
[d] *Ningbo Institute of Technology, Zhejiang University, Ningbo 315100, China*

Received 23 March 2017; accepted 24 March 2017
Available online 27 March 2017
Communicated by L. Perlovsky

Abstract

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. Since this constraint of memory is common for all human beings, there may well be a universal preference for dependency distance minimization (DDM) for the sake of reducing memory burden. This human-driven language universal is supported by big data analyses of various corpora that consistently report shorter overall dependency distance in natural languages than in artificial random languages and long-tailed distributions featuring a majority of short dependencies and a minority of long ones. Human languages, as complex systems, seem to have evolved to come up with diverse syntactic patterns under the universal pressure for dependency distance minimization. However, there always exist a small number of long-distance dependencies in natural languages, which may reflect some other biological or functional constraints. Language system may adapt itself to these sporadic long-distance dependencies. It is these universal constraints that have shaped such a rich diversity of syntactic patterns in human languages.

Keywords: Dependency distance; Language universal; Syntactic patterns

1. Dependency Distance Minimization: a human-driven linguistic universal

The founder of modern linguistics, Saussure [1], held linearity as one defining feature of human languages: a sentence is always produced or received linearly, one word after another. However, underlying the apparent linearity of a sentence is a hierarchical syntactic structure that closely bears on both production and comprehension. According to Dependency Grammar [2–5], this structure is largely composed of dependency relations between a word and another that syntactically governs or depends on it (Fig. 1).

* Corresponding authors at: C. Xu, School of Foreign Studies, Anhui Jianzhu University, Hefei 230601, China; J. Liang, Department of Linguistics, Zhejiang University, Hangzhou, 310058, China.
E-mail addresses: adinxu@126.com (C. Xu), jyleung@zju.edu.cn (J. Liang).

http://dx.doi.org/10.1016/j.plrev.2017.03.002
1571-0645/

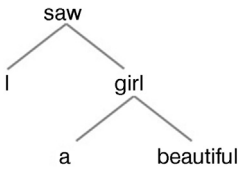

Fig. 1. The hierarchical dependency tree of a sentence.

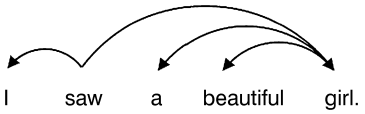

Fig. 2. The linearization of a hierarchical syntactic tree.

When such a hierarchical structure is linearized into the word sequence of a sentence, the dependencies may be adjacent or nonadjacent, that is, the two words forming a dependency may neighbor each other, or be separated by other intervening words, as shown in Fig. 2, and hence arises the concept of dependency distance or dependency length, defined as the number of words intervening between two syntactically related words, or their linear position difference in sentence. Recently, a succession of empirical investigations have reported a tendency for various languages to reduce dependency distance or dependency length of sentences [6,7], generally considered as shaped by the principle of least effort [8].

Dependency distance and dependency length are basically the same, both reflecting how far away a word is separated from another one that depends on or governs it. However, *length* seems to imply the result of measurement of an existing entity, carrying a strong static flavor. Hence *distance* is preferred in this review to suggest a dynamic process. For example, it is the distance, not the length, which constantly varies between a person and his destination until he stops moving and stands still. Similarly, in the process of both language production and comprehension, the distance of a dependency is unknown until both words are settled to establish this dependency. In both language production and comprehension, the brain, equipped with knowledge of language systems, operates constantly to construct dependency structures, or various dependency relations at various levels. Dependency distance grows and varies with the dynamic process of building a dependency, and gets fixed when a dependency link is finally established. So, in comparison with dependency length, dependency distance is probably a better term that can reflect language processing as a dynamic cognitive and psychological process.

In addition, these two terms are often associated with two different ways of deriving their first central moment (the mean) from large-scale language materials. Liu [9] uses the term Dependency Distance, and calculates the mean dependency distance (MDD) of a sentence or a text, using the following two formulas:

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

where 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, and

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

where n is the total number of words in the text, s is the total number of sentences in the text. Hudson [5] and Ferrer-i-Cancho [10] use more or less similar ways to calculate the MDD of sentences.

Temperley, Futrell and some other researchers [7,11–13] use the term Dependency Length (DL), taking the mean of total dependency length of sentences (MDL) as the metric. They first calculate the total dependency length of each sentence, and then figure out the average of these totals. The problem is that MDL is rather sensitive to sentence length, and therefore, the difference in MDL may be the result of factors like text genres, and cannot reliably indicate difference in syntactic difficulty. Lengthening a sentence (increasing the number of words) will surely increase its DL. However, if the increase of sentence length does not lead to long dependency relations, there may well be, according

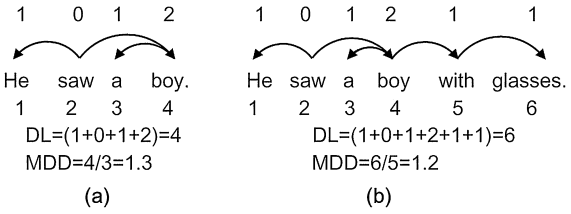

Fig. 3. Different influence of sentence length on DL and MDD. The dependency distance of each word is marked on the top, with the DD of the predicate verb being 0.

to dependency parsing, no increase of processing cost, which can be captured by MDD of this sentence, not the MDL, as seen in Fig. 3. This makes another reason that Dependency Distance is preferred in this review.

As illustrated in Fig. 1 and Fig. 2, dependency distance is closely related to both hierarchical structures and linear orders of sentences, or rather the syntactic patterns of sentences. Interestingly, natural languages widely present a tendency toward dependency distance minimization (DDM), that is, a propensity to syntactically structure sentences in such a way so as to minimize its overall dependency distance. This tendency in syntactic structures is not a recent finding. In fact, it has been observed, in one form or another, for a long period of time, and many theories have been proposed to account for it. The first observation of syntactic preference for short dependency distance perhaps comes from Behaghel [14], who, of course, did not use the term *distance* then, and made no efforts to further trace the possible biological or cognitive motivations for this phenomenon. Yngve's [15] depth hypothesis, based on the processing models of Phrase Structure Grammar (PSG), is probably the first model that predicts, on the basis of cognition, a general syntactic preference for short distance. According to this hypothesis, more embedding depths (unclosed phrases) mean more intermediate non-terminal nodes stacked temporarily in memory during sentence production, which cognitively increases processing difficulty. Generally speaking, more embedding depths usually correspond to a longer phrase, more intervening words, and heavier memory burden. In this sense, this model indirectly correlates long dependency distance with heavy processing difficulty. However, what matters in the PSG-based processing models is the number of unclosed phrase nodes while distance itself is not taken into consideration, and is reasonably unmentioned in Yngve's hypothesis. Similarly, Frazier's principle of Late Closure [16] and Gibson's Property of Recency Preference [17], though bearing more closely on dependency distance, makes no direct mention of such concept as distance since these models are also grounded in PSG. The term *dependency distance* (G: Abstand), coined by Heringer and his colleagues, first appeared in their 1980 work on syntax [18]. Fifteen years later, Hudson [19] gave, from the perspective of dependency parsing, the first explicit definition of dependency distance — the number of words intervening between a governor and its dependent, and presented, in view of parsing models of DG, a cognitive formulation of the memory burden imposed by dependency distance on language processing. The DG-based parsing aims to establish syntactic relation directly between the word being processed and another one being stored in working memory that decays with time, which is held as one possible root of short-term memory forgetting [20,21]. Hence a longer distance between them means more memory decay and more processing difficulty because more energy should be committed to the activation of the stored word.

Interestingly, some PSG based models also adopt similar concepts as an index of processing difficulty. The most well-known are EIC (MiD) principle [22,23] and Dependency Locality Theory (DLT) [24,25]. Both of them relate processing difficulty with the number of intervening words.

The EIC (MiD) principle is proposed on the theoretical ground somewhat similar to that of Depth Hypothesis: limited memory capacity determines, to a considerable degree, syntactic difficulty. This calls for small recognition domain (RD), because the bigger the size, the more words that have to be kept in memory before a phrase node is built. Therefore there is an urge for early appearance of the immediate constituent (EIC) to minimize domains (MiD), which is confirmed by statistical studies of corpus.

Another model is DLT, also holding time-invoked decay as responsible for the processing cost, which is termed as the locality effect — the distance-invoked difficulty. In this sense, Hudson's model of dependency distance can be seen as a dependency version of DLT, but capable of convenient data-processing in sentences annotated by dependency grammar. One major difference between them is that, in DLT, only new discourse referents are counted as the metric of decay while in Hudson's theory all intervening words are counted. The locality effect has been confirmed

in psychologically experiments: more discourse referents do invoke more processing cost, as reflected by reading time.

Given the principle of least effort, these different models and theories seem to suggest the universality of the tendency toward dependency distance minimization in languages. This tendency is shaped by external constraints, especially that of the limited working memory. For many years, the quest for linguistic universals has been an enduring endeavor of linguists, as can be seen in Speculative Grammar [26], Port-Royal Grammar [27], Notional Category [28], Universal Grammar [29], and Linguistic Typology [30,31]. These theories largely seek linguistic universals on the basis of philosophy, logic, or linguistic system itself, often neglecting the role of the common biological and psychological basis of language use. That is, the role of human beings is much overlooked. Even Chomsky's universal grammar, which is claimed to reflect the language faculty of brain, is not firmly grounded on established cognitive theories. Due to the overlook of use and users of languages, it is not surprising for Ferrer-i-Cancho [32] to contend that empirical means like statistics does not play a central role in the work of Saussure and many of his successors. Empirical investigations into parole or performance are of less importance than reasoning and speculation in seeking linguistic universals, with linguistic typology probably being an exception. This disregard for psychological and biological processes underlying languages and empirical studies of language use often lead to linguistic universals that contradict with empirical linguistic data, or fail to provide enlightening explanations, or both.

However, Dependency Distance Minimization (DDM) is probably a linguistic universal somewhat different, whose studies reflect a shift from the traditional rationalism and structuralism to a view that language is a complex adaptive system. In light of theories of complex adaptive system [33–35], language may be interpreted as a dynamic system, not self-contained, but shaped under the multiple external constraints like communicative functions and cognitive mechanisms that underlie the comprehension and production of language (in fact communication is ultimately cognitive or biological processes). In other words, it is human beings who use languages, and languages in use will adapt to, through self-organization, the communicative needs and cognitive constraints of human beings. Therefore, linguistic universals, like the tendency toward DDM, may largely be reflections or embodiments of deep-level human universals — be they functional, psychological, biological or even physical. In short, language is a human-driven complex symbol system, and linguistic universals, which may well be human driven as well, are some general patterns or tendencies formed as a complex system of language that adapts itself to multiple common human constraints [36]. Dependency Distance Minimization is probably such a linguistic universal.

Of course, human constraints only take effect when language is used by human beings. Therefore, human-driven linguistic universal may well be sought by examining the use of real language — the human verbal behaviors, not through speculating about highly abstract rules of langue or competence. Empirical means, such as psychological experiments, thus are playing an important role in detecting the patterns of language use. In addition, due to the development of computer technology, various corpora are now within easy reach of researchers, enabling them to statistically uncover, through big data analysis into vast recorded verbal behaviors, language universals, or rather, general linguistic tendencies that are hidden beneath the diversity of languages. As a matter of fact, the general tendency toward DDM is largely investigated through big-data analysis and computational simulation. Since DDM is generally considered as a linguistic universal shaped by common psychological and biological constraints to ensure efficient communication, the investigations into it may not only help us understand how language, as a complex system, operates and coevolves with human beings, but also boost our understanding of the mechanisms and laws of cognition and biology.

In the following sections, we will present various empirical data that support DDM as a universal of language, discuss the various syntactic patterns that contribute to DDM, and explore the possible mechanisms responsible for the sporadic long-distance dependencies that arise, against the tendency toward DDM, in natural languages.

2. Dependency Distance Minimization: psychological experiments

As an alleged linguistic universal, DDM is proposed on the theoretical basis that it is the external constraints of working memory that make languages evolve into such syntactic patterns to reduce dependency distance. To support this hypothesis, empirical evidences have to be obtained through close examination of how languages are actually used. The earliest evidences came probably from comprehension experiments on different types of relative clauses (RC) [24,37–40]. In English, for example, longer reading times have been recorded at the verb of an English object-extracted (OE) (RC) — such as "*the reporter that the senator attacked*" — than those at the verb of a subject-

extracted (SE) RC — such as "*the reporter that attacked the senator*", probably a result of the fact that in an English SE RC, the dependency distance is shorter between the verb of the RC and the relative pronoun than in the OE RC [41]. In Chinese, where RCs precede head nouns, the distance between the verb of RC and the head noun is shorter in OE RCs than in SE RCs, and accordingly, experiments found shorter reading time at verbs of Chinese OE RCs than SE RCs [42,43]. In German, the relative ease in processing short dependencies is also found, in some cases, in subject-extracted relative clauses [44]. Similar ease is found in a study of Russian relative clauses that controlled extraction type and word order [39]. In Russian the word order is rather free since it is case-marked. For example, "*Slesar'* (*repairman*) *kotoryj* (*who*) *udaril* (*hit*) *elektrika* (*electrician*$_{accusative}$)" and "*Slesar'* (*repairman*) *kotoryj* (*who*) *elektrika* (*electrician*$_{accusative}$) *udaril* (*hit*)" are both acceptable Russian SE RCs with virtually the same meaning: "*the repairman who hit the electrician*". Similarly, "*Slesar'* (*repairman*) *kotorogo* (*whom*) *elektrik* (*electrician*$_{nominative}$) *udaril* (*hit*)" and "*Slesar'* (*repairman*) *kotorogo* (*whom*) *udaril* (*hit*) *elektrik* (*electrician*$_{nominative}$)" are both acceptable OE RCs with the same meaning: "*the repairman whom the electrician hit*". Regardless of the extraction type, processing ease is invariably found in the cases where the verb follows the relative pronoun [39].

In other syntactic structures, recent psychological studies have also found supportive evidences [45–47]. In English, reduced dependency distance was reported to have a facilitating effect on the processing of subject–verb dependencies in main clauses, even when the materials are quite simple [45]. The acceptability of multiple wh-phrase (interrogative) constructions also depends, to a considerable degree, on the distance between the filler and the gap [48,49]. In Hindi Noun–Verb complex predicate, prolonged distance between the noun and the verb is found to bring about more reading time when expectation is not strong [50]. In Chinese, longer distance between the verb and the object also leads to increased reading time at the object [51]. Other evidences for the preference for short dependency distance come from the phenomenon of local interference, which is somewhat related to Frazier's late closure [52]. For example, in the structure *The key to the cabinets was* . . . , though the verb *was* is the governor of *the key* and agrees with it, it is not rare for readers to be interfered by the adjacent plural noun *the cabinets* and promptly relate it to the immediately following verb [53,54].

These experiments on various languages seem to point to a law that the processing load usually increases with the distance of dependency. In view of the least effort principle, this may lead to a wide preference across different languages for short dependencies. Admittedly, psychological experiments play vital roles in tracing general linguistic regularities and their underlying motivations — this approach enables researchers to control variables so as to probe into potential relations between independent and dependent variables in various settings. However, psychological experiments usually feature laborious design and high cost, which more than often confine studies to a small number of subjects and a rather limited range of artificial linguistic materials of a very limited number of languages. These limitations sometimes may lead to conflicting results, as can be seen in the studies of Chinese relative clauses: some [37,40,42,43] suggest that ORCs are easier to process than SRCs, while others [55–57] find the reverse. Therefore, when it comes to language universals like the preference for short dependency distance, corpus-based quantitative study, usually focusing on general tendencies in order to draw a macro picture of language, may serve as a significant supplement to psychological experiments, and hence have enjoyed growing popularity in linguistic studies. As one kind of human behavior, verbal communication is regulated and restrained, to a considerable degree, by some deep-level psychological or biological constraints, and hence may exhibit some human-driven patterns or regularities. These deep-level mechanisms are likely to be similar across different peoples. Therefore these patterns or regularities of language use are probably universal among users of different languages, making some linguistic universals hidden in the parole or the performance. With the development of computer technology, big-data based statistical analysis has been playing an increasingly important role in detecting patterns in various human behaviors [58]. A large-scale corpus, which provides an easy access to big data of verbal behaviors, may contribute significantly to scientific linguistic researches purposing to detect hidden linguistic patterns and to trace their motivations. To recapitulate, if Dependency Distance Minimization (DDM), or preference for syntactic structure with short dependency distance, is a general human-driven tendency in language, it may well be manifested and captured in big-data analysis of corpus, especially the dependency treebanks (i.e. the corpora annotated with dependency schemes), which explicitly annotate syntactic (dependency) relation between the governing words and the dependent words in sentences. In contrast, dependency relations are not explicitly marked in Phrase Structure treebanks, which renders it rather difficult to quantitatively investigate dependency distance.

3. Dependency Distance Minimization: corpus-based investigations

In contrast with psychological experiments, which are mainly concerned with dependency distance in specific syntactic structures, big data analysis of dependency treebanks often statistically investigates whether languages present an overall tendency toward short dependency distance. A corpus-based study of Romanian and Czech that controlled sentence length compared natural sentences with **random** sentences (sentences with scrambled words but intact dependency structures), pointing out that the average dependency distance of natural sentences is significantly shorter than that of random sentences, which may bear on another finding that in random sentences, the average dependency distance is a linear function of sentence length: $\langle d\rangle = (n+1)/3$, while in real sentences, the fitting function is $\langle d\rangle = 1.163 \pm 0.039n$, which suggests a much slower growth rate of average dependency distance in natural sentences than in random sentences [10]. These two findings clearly point to a general tendency toward DDM in these two languages, which is absent in artificial random languages [10]. Different from this study, another investigation into Chinese did not control the sentence length, and, to eliminate possible influences of syntax and meaning, created random sentences by scrambling dependencies, not words: for each sentence one word is randomly chosen as the root, and then for every other word another word is randomly chosen as its governor. Despite these differences, similar findings were reached that the overall mean dependency distance (MDD) of Chinese, as calculated with formula (2), is significantly shorter than that of the two randomly generated languages [9]. In these studies, MDD is used as the index of overall dependency distance. Nevertheless, instead of MDD, Gildea and Temperley used the sum of dependency length in a sentence as the measure, and compared the average of all the sums (MDL) in corpora of natural languages with that of corresponding random languages [13]. Though adopting a quite different index for comparison, their investigations, once again, reported significantly shorter average total dependency distance in both German and English than that in random sentences [13]. So, despite the various indices and the different baseline random languages, these studies seem to repeatedly indicate a universal that natural languages tend toward DDM: the relatively short dependency distance is probably not a statistical contingence or artifact, but a recurring linguistic regularity of natural languages.

However, all these findings are based on only a handful of languages, and thus may not serve as compelling evidences for a hypothesized linguistic universal like DDM: to establish linguistic universals, large-scale cross-language studies are a must. Anyway, to be universal, a linguistic pattern or tendency should persist in various languages.

The first large-scale cross-language study, which investigated into dependency treebanks of 20 languages, yielded supporting results for the hypothesized universality of DDM: all the 20 languages exhibit the same regularity that the mean dependency distance is significantly shorter than those of the two corresponding random languages that were created in the same ways as adopted in Liu [6], with the longest MDD below 4, as can be seen in Fig. 4. Interestingly, for all the 20 languages, the random language 2 (RL2), which is constrained by projectivity (the prohibition of crossing dependencies, as illustrated in Fig. 5a), consistently presents significantly shorter MDD than random language 1 (RL1), which, as illustrated in Fig. 5b, is not constrained by projectivity [6]. This finding suggests that projectivity, as a property of most languages, may contribute much to the short MDD of natural languages. Similarly, researches based on other languages [59,60] report quite limited mean dependency distances of natural languages, never exceeding 4, which seems to be a threshold of MDD of natural languages. Recently, another cross-language study [7] was conducted into 37 languages, which controlled sentence length, adopted the average of sums of dependency distance as index of overall dependency distance, and created random languages simply by scrambling the words, not the dependencies. Despite these differences, the findings were similar: the overall dependency distance of all the 37 languages are invariably shorter than the random baselines, which is a further piece of evidence that dependency distance minimization is probably a universal regularity in human languages.

All these corpus-based studies clearly point to a universal tendency toward DDM in natural languages, which suggests some underlying laws in natural languages that might have shaped this universal tendency, especially those concerning the distribution of dependency distance, which bears directly on the overall dependency distance of a language. According to Ferrer-i-Cancho [10], the minimum entropy principle could elicit an exponential distribution of dependency distance:

$$P(d) = \alpha(n-d)e^{-\beta d}, \tag{3}$$

where α and β are parameters, d is dependency distance, and n is the sentence length as measured in terms of word tokens. This exponential distribution predicts decrease of dependency frequency with increase of dependency distance,

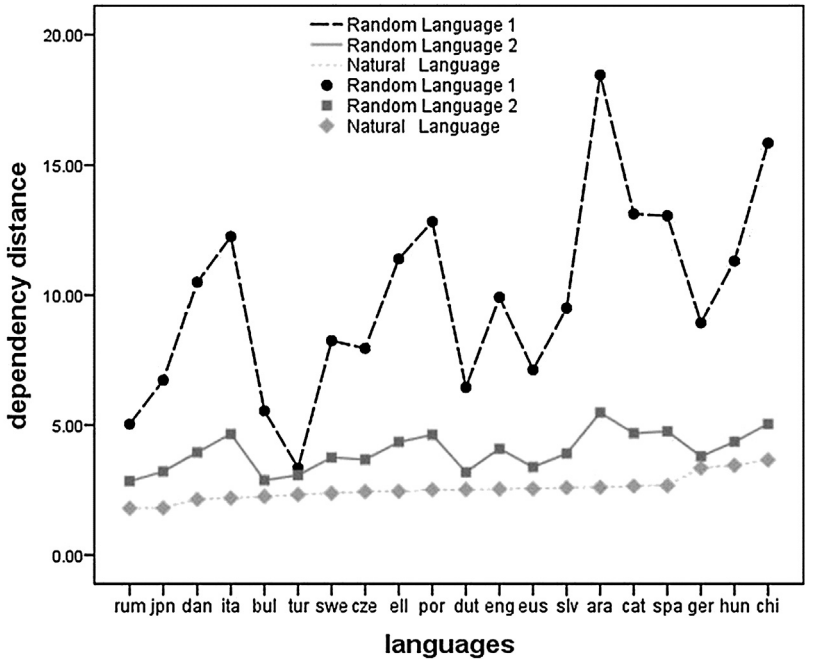

Fig. 4. MDD of 20 natural languages[1] and the corresponding random languages. The MDD of each language is invariably shorter than the MDDs of the two corresponding random languages. The MDD of random language 2, which is constrained by projectivity (that is, more similar to natural languages), is invariably shorter than the MDD of random language 1, which is not constrained by projectivity (that is, less similar to natural languages) (adapted from [6]).

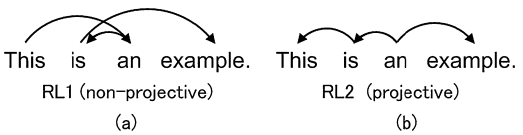

Fig. 5. Non-projective and projective random languages (RLs).

and was supported by empirical data, especially when the dependency distance is short [10]. However, some other studies seem to suggest a power-law distribution of dependency distance in natural languages, which also predicts that most dependencies are quite short and that the frequency of dependencies drops drastically with the increase of dependency distance. Based on a Chinese dependency Treebank, Liu [9] reported that in Chinese, the distribution of dependency distance, as shown in Fig. 6, can be captured by the right truncated Zeta distribution:

$$P_x = \frac{1}{x^a \left[\Phi\left(1, 0, a\right) - \Phi\left(1, R, a\right)\right]} \tag{4}$$

where x is dependency distance, α is a parameter, and R is the longest dependency distance in the text. This distribution is a power law distribution, which is absent in the baseline random language that was not constrained by projectivity.

A recent cross-language investigation, which is based on 30 languages, may partly settle this inconsistency [61]: the distribution of dependency distance is reported to follow different models — for long sentences, it seems to conform

[1] These 20 languages are: Arabic (ara), Basque (eus), Bulgarian (bul), Catalan (cat), Chinese (chi), Czech (cze), Danish (dan), Dutch (dut), English (eng), German (ger), Greek (ell), Hungarian (hun), Italian (ita), Japanese (jpn), Portuguese (por), Romanian (rum), Slovenian (slv), Spanish (spa), Swedish (swe), Turkish (tur). Language codes are following ISO 639-2.

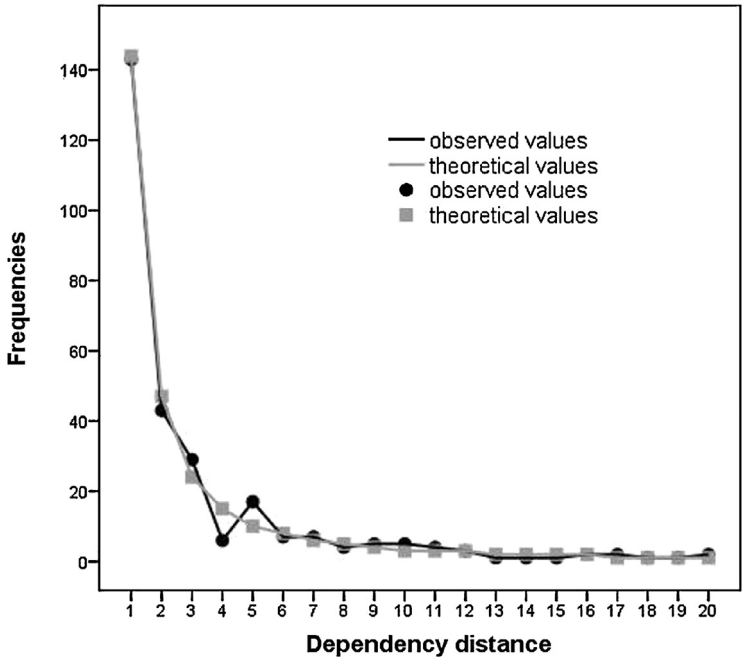

Fig. 6. Fitting the right truncated Zeta distribution to the dependency distances in a Chinese text. The observed data fit well with the expected values, with the frequency of dependencies dropping dramatically with increase of dependency distance.

to the power law distribution:

$$p(x) = \frac{\lambda^{1-\alpha}}{\Gamma(1-\alpha, \lambda x_{min})} \cdot x^{-\alpha} e^{-\lambda x}, \tag{5}$$

where Γ is the gamma function, λ is the rate parameter, α is a constant parameter of the distribution known as the exponent or scaling parameter, and x_{min} is the low bound. While for short sentences, dependency distance generally abides by the exponential distribution [61]:

$$p(x) = \beta\lambda e^{\lambda x^{\beta_{min}}} \cdot x^{\beta-1} e^{-\lambda x^{\beta}}, \tag{6}$$

where β is the shape parameter, λ is the scale parameter.

This may be due to the higher pressure in long sentences to curb dependency distance than in short sentences, because the principle of least effort is held as having shaped power law distributions of various linguistic measurements like frequency and length [8,62].

However, we perhaps needn't fuss over the specific distributions: both exponential distribution and power-law distribution suggest the dominance of short dependencies and the drastic decrease of frequency with growing dependency distance. One study fits six different power-law and exponential distributions to the dependency distances in a parallel English–Chinese corpus, only to find that the goodness of fitting is statistically acceptable for all these six distributions [63]. We also fitted the dependency distance in a Chinese treebank to some probability distributions used in Jiang and Liu [63], that is, Right truncated Kemp2, Right truncated Salvia–Bolinger, Right truncated Waring, Right truncated negative binomial, Right truncated modified Zipf–Alekseev [64,65]. Our study reached similar findings: goodness of fitting is statistically acceptable for all these distributions, as illustrated in Fig. 7. Another study reported that Right truncated Waring distribution, Right truncated Zeta distribution, and Exponential distribution can well capture the distribution of dependency distance in different genres of English, though with slightly different goodness of fitting [66]. These different distributions all present a similar regularity that the frequency of dependency drops drastically with the increase of dependency distance, with the number of the adjacent dependencies being the highest.

However, it is also argued that these distributions of syntactic dependency lengths could be a mere consequence of that mixing [67], because dependency distance may be susceptible to sentence length [10,68–70], which cast doubt on some previous studies [6,13] that, based on sentences of mixed lengths, reported not only significantly shorter MDD

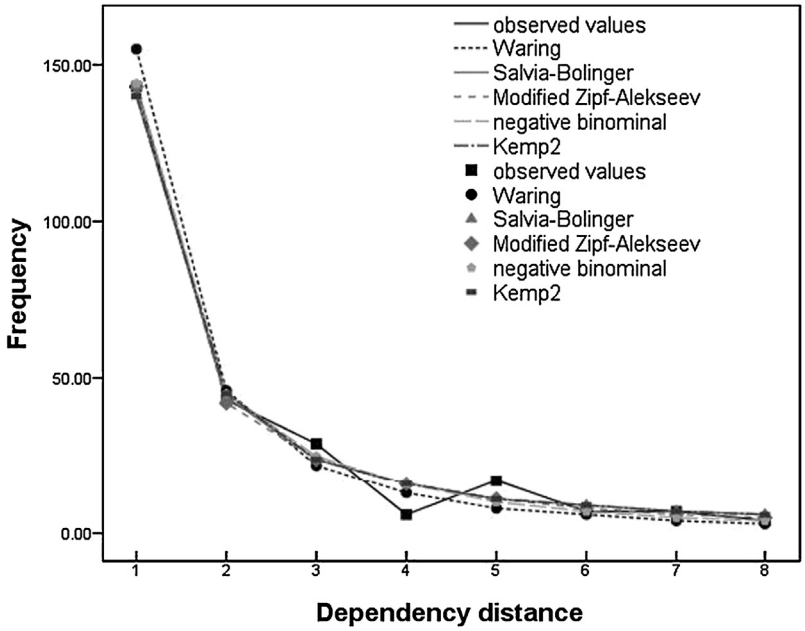

Fig. 7. Fitting of the DD in a Chinese Treebank to several distribution models.

of natural language than that of random languages, but also considerable variations in MDDs of different natural languages. These differences in the global MDD may merely be the result of improper sampling of sentences with mixed length, probably with no linguistic or cognitive significance [67].

To address this issue, a study investigated the distribution of dependency distance of sentences of various lengths, whose findings indicate that the long tail of distance distribution seems to persist in sentence sets of different sentence lengths [63]. These findings suggest that, given the relatively high proportion of adjacent or short dependencies, the mean dependency distance is likely to be highly constrained even in long sentences. This may be the result of common human biological mechanism, especially the working memory, which has a universal shaping effect on syntactic structures, restraining dependency distance of natural languages.

Therefore, the variations of MDD among different languages cannot be readily dismissed as contingent statistical artifact of careless sampling of sentences with mixed length, which therefore calls for further multi-disciplinary investigations. But more importantly, the long-tail distribution of dependency distance seems a fundamental linguistic law, be it a power-law distribution or an exponential one, and prescribes a persistent high proportion of adjacent dependencies and the virtual immunity of distance distributions to sentence length. This distribution of dependency distance suggests very powerful and effective mechanisms and patterns in syntactic organization that prevent MDD of long sentences from drastic rising. For example, Fig. 4 illustrates that the projectivity, a general pattern of languages that prohibits crossing dependencies, may considerably reduce the MDD of random sentences. These mechanisms and patterns probably result from gradual adaptation of language system to external constraints and pressures in evolution, and thus make a significant part of grammar, as embodied in recurrent syntactic patterns or laws that can be captured through big-data analysis.

4. Syntactic patterns and regularities shaped by the pressure for DDM

The artificial random languages used in all the above studies differ from natural languages in that they are hardly bound by syntax or grammar, created by scrambling the word order or the dependencies of natural languages. The only exception is the second random language used in Liu [6,9], which is syntactically constrained by projectivity, presenting significantly shorter MDD than those of other random languages. These facts lead to a logical assumption: the lack of grammatical or syntactic organization in random languages may be a major reason for the long overall dependency in random languages. In natural languages, in contrast, the cognitive constraint of limited working memory may have shaped many syntactic patterns, in the forms of syntactic regularities that are either traditionally assumed

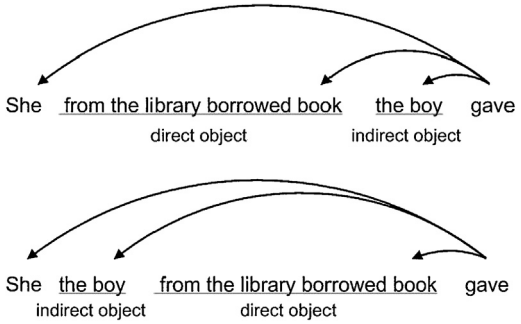

Fig. 8. Long-before-short and short-before-long arrangements in Japanese.

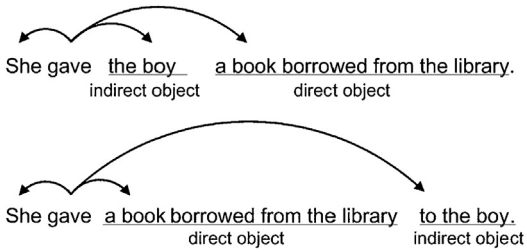

Fig. 9. Short-before-long and long-before-short arrangements in English.

as hard and absolute rules or statistically dominant options that are recently demonstrated 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. To recapitulate, the pressure for DDM may have driven a considerable number of syntactic patterns of natural languages.

Dependency distance is measured in terms of the linear positions of words in a sentence, suggesting that the dominance of short-distance dependencies may directly bear on the serial arrangements of words in a sentence, or, the syntactic patterns of word order. In various languages, word orders, despite their apparent diversity, seem to universally present a general regularity to make for DDM. At the phrase level, a recent study, which fitted a logistic regression model to corpus data of noun phrases, found the length of internal constituents as a significant predictor of the order of constituents — the internal constituents tend to be arranged by their relative lengths so as to minimize the overall dependency distance [71]. At the sentential level, similar regularities were noticed that sentential constituents are ordered according to their lengths [14], with the last constituent being the longest [72], which is termed as the principle of end-weight [73], alleged to dominate the ordering of post-verbal constituents. This principle is empirically supported by corpus-based findings that post-verbal behaviors, like heavy noun phrase shift, particle movement and dative alternations, depend substantially on the weight or the relative lengths of the constituents, usually leading to a distance-reducing short-before-long order [74,75]. Hoffmann [76] claimed that such a short-before-long linearization is cognitively motivated by the pressure to reduce the recognition domain, or, the distance between the verb and the head of the last constituent governed by the verb. Nevertheless, some availability-based accounts are held as able to better explain the weight effects than the distance-based accounts [77,78]. Interestingly, the corpus analyses [22] show that, Japanese, which is a SOV language, presents a recurring pattern of long-before-short arrangement: long preverbal phrases tend to occur ahead of short ones. Some on-line experiments have also confirmed this long-before-short preference among speakers of Japanese [79]. Similar patterns of long-before-short arrangements of preverbal constituents are found, through both corpus study and psychological experiments, in other head-final languages like Persian and Korean [80–82], which denies the explanation that the more available constituents come earlier in production. In a SOV language like Japanese, the long-before-short arrangement of preverbal constituents makes for short dependency distance, as can be seen in Fig. 8, while in SVO languages like English, it is the short-before-long order of post-verbal constituents that makes for short dependency distance, as can be seen in Fig. 9.

This contrast is interesting — the same cognitive pressure to reduce dependency distance may lead to different or even contrary patterns of word order, as seen in the long-before-short preference in SOV languages and the

short-before-long preference in SVO languages. However, in human languages these contrary patterns are no denying general linguistic laws, but open a window for us to ascertain the hidden universals and laws that give rise to the rich diversity of human languages.

Given that the DDM is a universal pressure motivated by deep-level constraints like the limited memory of human beings, it may well have wide influence upon word order, not merely on the specific syntactic structures discussed above. As a result, some universal and abstract linguistic laws [11,12], which are not confined to specific languages or syntactic structures, have been proposed as governing the sequential order of syntactic structure to ensure short dependency distance. These laws are generalized as Dependency Length Minimization Rules (DLMRs), including consistent branching, opposite-branching of one-word phrases, and nesting phrases. The potency of these principles has been confirmed by computational simulations, whose results indicate that artificial grammars framed on these rules yield substantially shorter dependency distance than artificial grammars that are not constrained by these rules [12]. For example, there is a DMLR that "if a word has multiple dependent constituents and there is a choice as to their ordering, the shorter one(s) should be placed closer to the parent head", and this rule, when integrated into an artificial grammar, may significantly restrain the dependency distance in the sequences generated by that grammar [12]. Corpus-based study also finds that in English and German, syntactic patterns conforming to these rules are generally preferred to those running counter to them [11].

However, it should be pointed out that these rules or laws reflect statistical regularities — some overall tendencies or dominant regularities that cannot be held as invariable and absolute in the linearization of words in sentences. In fact, the dependency distance of natural languages, though shorter than that of random languages, never reaches such minimum as predicted by DLMRs [7] and word orders of different languages seem to agree with DLMRs to different degrees [13]. Hence, language universals are probably more of general tendencies than absolute rules. In some cases, long dependencies may be acceptable or even preferable. As a complex system, a language operates and evolves under multiple constraints. One of them demands short dependency distance, while other constraints may compete or interact with it, leading to some occasional long dependencies. To recapitulate, there are probably other factors that interact with the pressure for short dependency distance, or even with the biological constraint of limited memory, to shape the rich inter-language diversities and intra-language variations. This represents a complex system view of language, probably best seen in synergetic linguistics [83], which approaches languages with the complex system theory and quantitative means, regarding language as a system governed by a few general constraints, competing and interacting with each other to mold various linguistic patterns emerging out of language evolution.

DLMRs are syntactic regularities in word order that contribute to DDM. However, the linear order of words is merely one syntactic aspect that bears on dependency distance. Other regularities, which pertain to other aspects of syntax, may also arise from the operation of language system to meet the requirement for short dependency distance. For example, the word order of SOV conduces less to the minimization of dependency distance than SVO order. This may be mended by a shift of word order from OV to VO, or by frequent omission of arguments in sentences, or both. As an SOV language, Japanese is theoretically expected to have longer dependency distance than SVO languages like English. But corpus studies indicate no significant distance difference between them, probably a result of another statistical regularity that there are significantly more omissions of dependents in Japanese than in English, which Hiranuma interpreted as a specific syntactic option to reduce the possible long distance caused by the SOV word order of Japanese [59]. Similarly, other corpus studies [84] also point out pro-drop and intransitive bias as two peculiar syntactic regularities in SOV languages, possibly arising from a similar pressure to reduce dependency distance. Again, the universal pressure of reducing dependency distance leads to considerable linguistic diversity, which is not surprising: slightly different initial conditions may drive a complex system to evolve into very different patterns even though external constraints are the same, and language is no exception. Different initial conditions of languages may lead the same constraint to shape different syntactic patterns. The apparent diversity of grammar may actually be the results of simple and common constraints, such as the memory constraint that demands short dependency distance. There are numerous patterns that may emerge from a complex system to meet an external constraint or requirement. It may be easy to explain how one pattern may meet a certain requirement, but difficult to predict which one or which ones will become dominant, since it is sensitive to even slight differences in initial conditions. This seems also true of language: the memory constraint, reflected by DDM, can be met through diverse syntactic patterns in a language system.

Optimizing word order or reducing arguments of a verb both contribute to short dependency distance. Beside these apparent syntactic regularities that directly influence the sentential representation, there are some other syntactic pat-

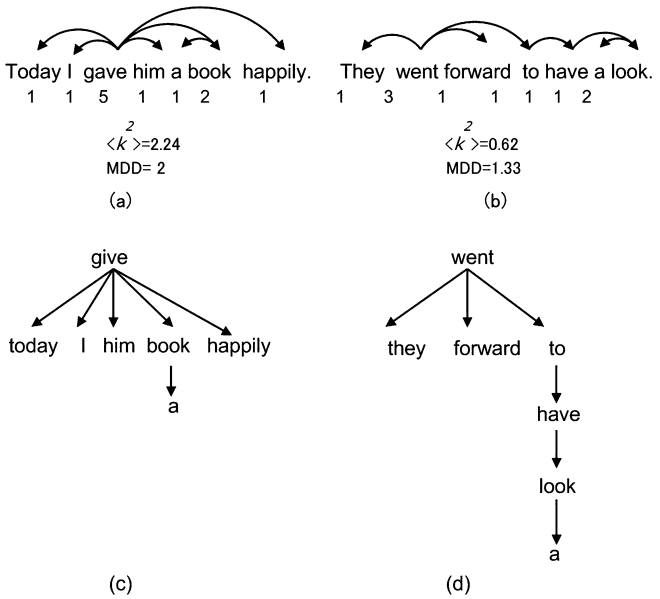

Fig. 10. A rough illustration of the relations among the degree centrality, hierarchical depth and mean dependency distance. The number under each word in (a) and (b) indicates the degree of each word. It can be seen that the sentence in (a) has higher $\langle k^2 \rangle$ and longer MDD than the sentence in (b), which means higher degree centrality, more parallel words and fewer hierarchical depths, as can be seen in (c) and (d).

terns less apparent. Every sentence has a syntactic structure hidden in the apparent linear sequence. In terms of either the Phrase Structure Grammar or Dependency Grammar, this syntactic structure is a hierarchical tree, and the pressure for short dependency distance plays a role in shaping the overall graphic features of these hierarchical syntactic trees. One graphic regularity in syntactic trees is the rarity of crossing dependencies in natural languages, which is supposed to be a side-effect of distance minimization [68,85,86]. A mathematical model, into which dependency distance is integrated as a feature, has been reported as capable of effectively predicting the probability of two dependencies crossing each other [87]. Corpus-based studies point to considerable validity of this distance-based model in estimating and predicting proportion of dependency crossings in real languages [88]. At the same time, another study based on computer enumeration similarly reveals a relation between dependency distance and dependency crossing: the manipulation of MDD has an effect on the number of dependency crossing in the dependency tree generated by computer algorithm [89]. According to these studies, the rarity of dependency crossings seems to be a sentential graphic regularity driven by the pressure for short dependency distance [6].

Apart from crossings, another graphic feature relevant to DDM seems to be degree centrality, which reflects a vertex's relative importance in terms of its degree, measured by the number of edges incident to this vertex in a graph, that is, the number of other vertices linking to it. In a network, a node with a much higher degree centrality than other node is a hub of the network. The distribution of degree centrality in a network or graph can be reflected by the second moment of degree $\langle k^2 \rangle$, which measures the variance of the degree. High $\langle k^2 \rangle$ means that the degree is distributed rather unevenly: a few nodes have quite high degree while other nodes have quite low degree. These nodes with high degree are hubs. In general, high degree centrality linguistically means that a head (or a few heads) takes many parallel dependents, and that the syntactic tree does not have much hierarchical levels. That is, the dependency tree is rather flat. So, when linearly arranged, some dependents inevitably will intervene between the head and other dependents. For a dependency tree, low degree centrality may contribute to DDM, as can be seen in Fig. 10.

One study indicates that, the minimal average dependency distance is bounded below by the variance of degree, that is, the larger the second moment of vertex degree is, the longer the minimum average dependency distance [85], which is mathematically supported by another study [90]. In other words, high hubiness should be avoided to reduce dependency distance, which linguistically means that a head should not take too many parallel dependents and that a hierarchical structure with many levels is preferred for the sake of DDM. The reason is simple: if a head takes many parallel dependents, when linearly arranged, some dependents inevitably will intervene between the head and other dependents, and thus leads to long dependency distance. When constrained by projectivity, for the linear tree (Fig. 7a),

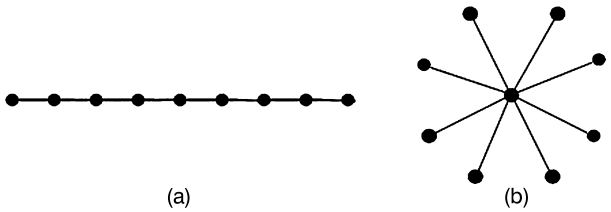

Fig. 11. A linear tree and a star tree.

the minimal average dependency distance is 1, and for the star tree (Fig. 7b), the minimal average dependency distance is $(n^2)/[4(n-1)]$ when the number of nodes (n) is even, and $(n+1)/4$ when the number is odd [85]. Intuitively, the linear tree has a better chance to yield shorter total or mean dependency distance than the star tree.

What is interesting is that, mathematically, the number of crossing dependencies is expected to be minimal in a star tree, which has the highest $\langle k^2 \rangle$ — the highest second moment of vertex degrees (reflecting degree centralization or the variance in degree distribution), and maximal in a linear tree, which has the lowest second moment of vertex degree [91], as can be seen in Fig. 11.

It has been mathematically alleged that, for random linear arrangements of a dependency tree, the expected number of crossing dependencies is a function of the number of the vertices of the tree and of the second moment of degree. That is, the higher the second moment of vertex degrees, the smaller the maximum number of crossings [85], as predicted by the following formula

$$C_{max} \leq \frac{n-1}{2}\left(n\langle d\rangle - \langle d^2\rangle - n + 1\right). \tag{7}$$

These results indicate a complicated situation. Considering the general tendency in natural languages to minimize dependency distance, the rarity of crossing dependencies is probably not due to high degree centrality (hubiness), but some unknown graphic features that may give rise to linguistic structure of projectivity and limit long-distance dependencies. In fact, the star tree only constitutes a very limited portion of dependency trees in natural languages, hardly found in long sentences.

In the two trees in Fig. 7, the links are all undirected. In syntactic trees, however, a dependency relation is always directed, indicating the unequal syntactic status of the two words. Hence, syntactic trees are hierarchical and, given a certain sentence length, lower hubiness usually involves more hierarchical levels. For example, a star tree has only 2 levels, while the linear tree in Fig. 7 has at least 4 levels. It then seems that syntactic trees with more hierarchical levels get a better chance to have short dependencies. Hierarchical trees, in their linear realization, are represented as chunks at different levels. Between any two chunks at one level, there is only one dependency, which means many intra-chunk dependencies and a handful of inter-chunk ones, thus contributing to reducing the overall dependency distance of a sentence. The contribution of chunking to DDM seems to be confirmed by a recent computational simulation [92], which found that chunking plays a significant role in reducing dependency distance, and that, as roughly shown in Fig. 12, in artificial sentences whose lengths are between 16 and 32, minimum MDD can be reached if the chunk size is set between 4–7, much close to the average clause length of natural languages.

Given that the hierarchical structure of syntactic trees usually corresponds to chunks in linearization of sentences, their study provides evidences that hierarchical structures may contribute to short dependency distance. In random languages, MDD is a linear function of sentence length, while in natural languages it only slightly increases with the growth of sentence length. Therefore, a graphic regularity of natural languages may be assumed that a longer natural sentence should have less degree centrality to avoid mushrooming long dependencies. That is, in a longer sentence, the sentence structure usually has more hierarchical levels (more chunks at more levels) to avoid too many parallel dependents. In contrast, shorter sentences may tend to have higher degree centrality and less hierarchical levels since parallel dependents would not bring about long dependencies when sentences themselves are short. For example, the degree centrality is quite high in the sentence "*yesterday, I gave him a present*", which, however, will not lead to many long dependencies since this sentence is quite short. This hypothesized regularity seems to be evidenced by the findings that texts with shorter sentences tend to have larger degree centralities [93], and that hierarchical depth increases with sentence length in natural languages [94], which is captured in another study by the fitting function $T = 1.8188P^{3.51}e^{0.00423P}$ [95]. The relation between hierarchical depth and sentence length can be roughly visualized

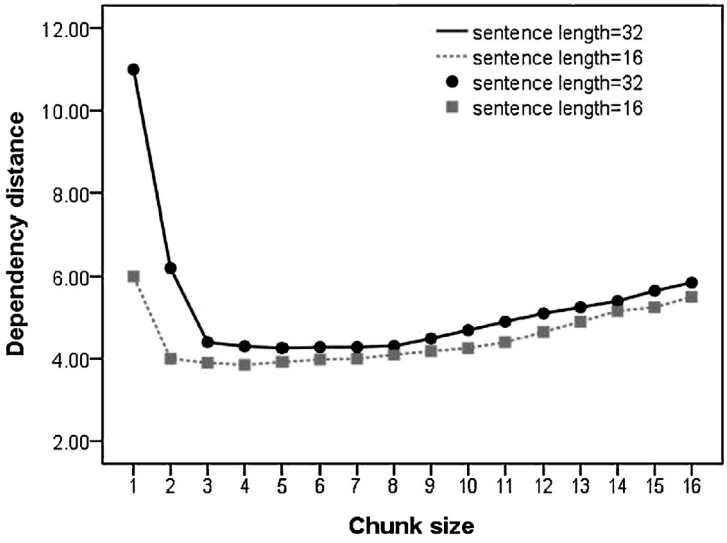

Fig. 12. The relation between the chunk size and the mean dependency distance of artificial sentences.

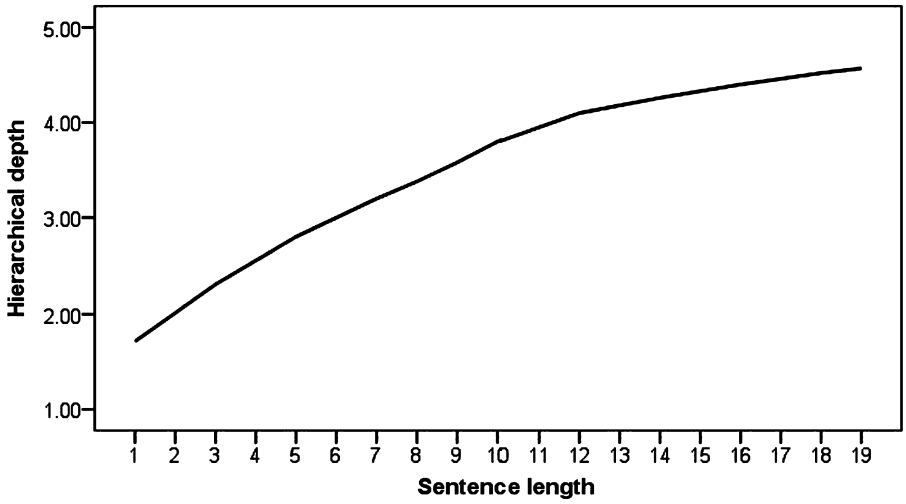

Fig. 13. The relation between hierarchical depth and sentence length.

as the curve in Fig. 13. In fact, it has been pointed out that mere projectivity fails to reduce the MDD long artificial sentences to that of natural language: only when combined by chunking (hierarchical organization) can it reduce the MDD of long artificial sentences to that of Mandarin Chinese [92]. These researches all reveal that the pressure for short dependency distance may have shaped some graphic patterns in syntactic trees, each playing its own role in reducing MDD in sentences of different lengths.

The pressure to reduce dependency distance, as revealed by these studies, may be responsible for various syntactic regularities in either linear realizations or graphic patterns, which are statistically dominant in the use of natural languages. These synchronic regularities are the results of diachronic shift, which implies that DDM may have also played an important role in shaping the evolution of languages — many diachronic variations of syntax may regularly present a general tendency toward short dependency distance.

To a considerable degree, the gradual change of word order is probably motivated by cognitive constraints to exhibit a general tendency toward reducing dependency distance. The Subject-Object-Verb (SOV) order of sentences, as found in "∗*I him saw*", is usually held as dominating early stages of human languages [96] — a hypothesis supported by computer simulations [97]. Gesture communication experiments, requiring participants to communicate

simple transitive events [98], report a similar universal bias for SOV order in simple communications. However, one gesture study has found that the possible confusion of subject and object may render a switch from SOV to SVO (Subject-Verb-Object) order [99], which suggests confusion or interference as a possible reason for the adoption of SVO order. In fact, interference from interveners is held as one possible reason why short dependency distance is preferred. Intuitively, complex sentences with long dependencies are likely to suffer from such confusion or interference, for the more intervening words there are between a governing word and a dependent word, the more likely they are to cause interference. Some studies do point out that the bias for head-final arrangement tends to disappear in complex sequences [98,100] — for complex sequences, a final head often involves many intervening words between it and its dependent, while a middle head may significantly reduce the distance between them. Therefore, with sentences evolving into increasing complexity and length [101], the constraint of memory might well have driven language to drift from the SOV order to SVO order, for the sake of limiting dependency distance. Based on the data of diachronic English corpus, a multilevel logistic regression model indicates that both the object length and the date are significant predictors of whether the OV or VO order is preferred, which means that, for long objects, a VO order was likely to be preferred even in old English and middle English which feature frequent OV construction, and that the diachronic change presented a trend for the VO order to prevail over the OV order [102]. For example, in Old English, a longer object tended to appear after the verb, making an SVO sentence like "*unbearable itching overcame all the body*", while a short object often occur between the subject and the verb, producing an SOV sentence like "*but he must the quarrelsome reconcile*" [102]. Ferrer-i-Cancho's interpretation of this phenomenon is that word order variation has two attractors: maximum predictability of the head and minimum online memory, and that SOV placement probably prevails in simple sequences because maximizing predictability may have the priority in short sequences, while the SVO order may thrive in long sequences because minimization of on-line memory cost may become more pressing in this case and thus play a key role in molding word order [100]. In other words, the SVO order is possibly one of those syntactic regularities motivated by DDM in the process of language evolution to adapt to increasingly long and complex sentences. It has been mathematically proven that, given that processing cost is a monotonically increasing function of dependency distance, the adaptive landscape defined by online memory cost has a form of quasi-convex, which means the attractor of central verb placement [103]. In other, the SVO order seems an optimal word order when DDM is exclusively taken into account.

This mathematical conclusion is supported by computational simulations, which capture the gradual diachronic drift from the OV order to the VO order in English by adding processing bias to a model of multi-cue variational learner [102]. In contrast, the computational simulation will yield an SOV order if it is based on the neural networks biased toward predictability [97]. So language evolution, which usually features increasing sentence length and complexity, may gradually turn up the pressure for minimum online memory cost and give rise to increasing SVO structures in order to limit the growth of dependency distance.

However, this does not mean that SOV languages necessarily involve longer dependency distance than the SVO order. In fact, though it has been suggested that head-final languages may have longer dependency distances, corpus studies of various languages give no evidences that head-final languages consistently present longer distance than head-initial one [6,104]. There may be some other syntactic patterns, in addition to word order, that emerge from evolution to limit dependency distance. For example, Japanese seems to have resorted to frequent argument omission to cope with the possible long dependency distance caused by the SOV order [59], which is evidenced by the corpus findings that intransitive clauses are often favored over transitive ones in SOV languages [84].

Apart from the evolution of word order and number of arguments, the pressure for short dependency distance may have shaped some patterns regarding whether a language evolves along an agglutinating or an inflectional dimension. To reduce dependency distance, sentence markers that indicate interrogation, negation, etc., are often placed before verbs in VO languages, and after verbs in OV languages, which gradually lead to a regular pattern of agglutinative structures in OV languages, and inflectional structures in VO languages [105].

Synchronically and diachronically, the constraint of memory probably makes dependency distance minimization an important part of grammar, shaping numerous syntactic patterns. Hence, in computational linguistics, which aims to model human capacity in processing natural languages, the introduction of dependency distance as a constraint into computational models has brought about significant improvements in both automatic parsing and generating [106,107]. In another study, the computational model was armed with two more refined constraints of activation decay and interveners' interference, both alleged as cognitively responsible for the processing difficulty of dependency distance, and simulation findings indicate that these two distance-related constraints enable the model to accurately predict

syntactic complexity in such linguistic phenomena as weak islands and superiority [108]. In short, the cognitively motivated tendency toward DDM is probably a linguistic law, a significant part in syntax and syntactic theory [109], and thus may serve as an important constraint in models of computational linguistics.

What is noteworthy is that the general tendency toward DDM is largely detected through statistical investigations into corpora and reflected by statistically dominant linguistic patterns or regularities, which implies that exceptions, i.e. long-distance dependencies, are inevitable. Corpus studies point out that the dependency distance of natural language, though significantly shorter than that of random languages, never reaches the theoretical minimum predicted by the DLMR rules [7]. Computational simulation also reports that the unqualified bias for dependency distance minimization often reduces dependency distance to such a degree so as to deviate from real languages, implying the necessity of some counter-balancing features or biases in the modeling of natural languages [106]. Diachronically, despite the fact that the mean dependency distance decreases in most cases [102,104], the quantitative study of Mandarin Chinese seems to report slight increase of dependency distance [110]. These findings suggest the existence of other shaping forces that interact or compete with the pressure for DDM to give rise to occasional long-distance dependencies in natural languages. Of course, these long-distance dependencies account for only a very small portion in real languages, and could lead to processing difficulty. However, it is also possible that some other regularities or patterns may arise to adapt to these inevitable long dependencies so that they don't always impose unacceptable processing load on human beings — language is a dynamic adaptive system. In other words, long dependency distance may motivate some syntactic patterns and regularities that can reduce the processing difficulty of actually used long dependencies to meet the constraint of memory and the principle of least effort. [99,111], which will be extensively discussed in the following part.

5. Syntactic patterns involving long dependency distance

Long-distance dependencies, though quite limited in number, seem unavoidable in natural languages [7,9,13], which may be attributed to other pressures to meet other requirements, such as reliability, coherence, etc. As a complex system, language is capable of self-organization and self-adaption, and, when dependency distance minimization has to be sacrificed for the sake of reliable and effective communication, it may exploit other strategies to counter-balance the memory burden imposed by long dependency distance and thus give rise to some unique linguistic patterns for long dependencies.

Limited working memory is widely held as the cognitive basis of distance-invoked processing difficulty. It was once assumed that limited memory capacity determined syntactic difficulty [15,22,112]. Recently it has become widely accepted that distance-related difficulty may be largely attributed to the time-decay or the interference of working memory, not simply the size of memory capacity.

According to some linguistic models, the activation of a word or a syntactic representation in memory decays as a function of time [20,21]. Therefore, longer distance between a head word and its dependent demands more energy in processing to reactivate the head (or dependent). Gibson's Dependency Locality Theory (DLT) contends that the time-invoked decay of activation leads to distance-invoked difficulty, or, the locality effect [24,25]. Word Grammar [5, 113] even argues that memory capacity is limited only because energy available for activation is limited. Therefore, it is assumed that code switching often occurs in long-distance dependencies, because longer dependency distance may reduce, as a result of time-decay, the influence of a word on its forthcoming dependent (or head) in a sentence [60], which seems to be supported by the corpus-based findings that monolingual dependencies generally present shorter distance than such Chinese–English mixed dependencies as that between "kanjian (seen)" and "book" in "*ta(he) kanjian* (*seen*) *le* (*perfective marker*) *yi(a) ge* (*quantifier*) *good book*" [114].

Then it can be assumed that, in a long dependency, if the intervening words can be quickly processed, the previous word may suffer only limited decay. This assumption seems to be supported by repeated findings that highly accessible intervening words may facilitate the processing of long-distance dependencies. For instance, as interveners, high-frequency word combinations, such as *I met*, can significantly facilitate dependency processing, which is absent in low-frequency word combinations like *I distrust* [115]. Other studies report similar findings that more accessible intervener like pronouns often make for easier processing of dependencies [24,48,49,108]. Thus there might be a linguistic regularity that long dependency distance may often involve highly accessible interveners.

Similarly, proper chunking may reduce the difficulty of long distance dependencies, since chunking is an important way to improve processing efficiency and reduce processing time [109]. Corpus-based studies of Chinese and English

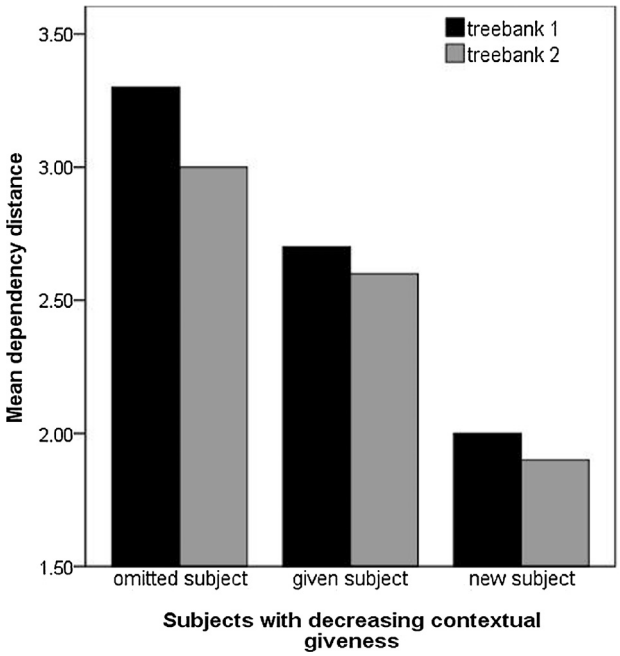

Fig. 14. Mean dependency distances of Chinese subjects with different degrees of contextual giveness.

seem to suggest a linguistic regularity that functional elements like particles or complementizer *that* are more likely to be used in long dependencies than in short ones [116,117]. 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.

Interveners in a dependency have considerable influence on the processing of the dependency, which, however, means no denial of the role of the two words that form this dependency. Some studies suggest that language processing may be affected by the segmentation cues of words [118], or their salient positions, especially the string-initial and the string-final positions [119,120]. For example, the experiments on artificial grammar learning show that only when the crucial syllables appear in such positions can artificial grammar patterns be learnt by participants [120]. This facilitating role of contextual cues is closely related with the recency effect and the primacy effect, which predict that, in a list of items, the ones most likely to be remembered are the items at the beginning and the end of the list [121]. In short, contextual cues, which usually lead to positional salience, may render a word less vulnerable to distance-invoked decay. Therefore, there might be another syntactic pattern that long dependencies often involve words at peripheral positions.

The role of positional salience actually suggests that high accessibility of a word in memory may allow a distant dependent or governor. Gibson holds a predicate verb as immune to dependency distance on the ground that it is the kernel of a sentence, and thus remains highly accessible in sentence processing [24]. Another study based on two Chinese treebanks suggests that the distance of Chinese subject dependency seems to be correlated with the degree of contextual givenness (Fig. 14) — probably higher contextual givenness renders a word more accessible in memory, and thus less insensitive to decay [111]. For example, in Mandarin Chinese, a contextually given (or familiar) subject, say "they", is more likely to be long separated from its predicate verb, than a contextually new (or unfamiliar) subject, say "clever students". In other words, the sentence "*tamen* (*they*) *zai* (*on*) *zhege* (*this*) *zhongyao* (*important*) *wenti* (*issue*) *will agree with us*" is more likely to occur than the sentence "*clever students on this important issue will agree with us*" in the actual use of Mandarin Chinese. In addition, since repeated reactivation of that word may increase its accessibility, a long dependency is not necessarily difficult to process if interveners themselves depend on or govern the previous word, and reactivate it time and time again, which can be formalized in an ACT-based language model [122,123].

Given these findings, it may be further assumed that other factors like frequency and conditional probability may also reduce the processing difficulty of long dependencies: high frequency or probability also contributes to high ac-

cessibility. It has been suggested in one study that the extremely high frequency might partly account for the immunity of Chinese preposition "zai (on, at)" to long dependency distance [124], which implies a possible regularity that words with higher frequency may allow longer dependency distance. Probabilistic Valency Pattern (PVP) [125] presents another possibility: the accessibility of word may be also affected by the conditional probability $P(A \leftarrow B \mid B)$ or $P(A \rightarrow B \mid B)$ — the probability of a previous word A depending on or governing B given that B appears. According to the theory of spreading activation, high conditional probability may suggest that, when a word is activated, the other receives considerable activations as well, and thus becomes easily accessible.

In short, the small number of unavoidable long-distance dependencies in natural languages are not necessarily difficult to process because some linguistic regularities and patterns may appear to take advantage of contextual familiarity, frequency, valency probability, positional salience, easy interveners, chunking, etc., to mitigate the decay caused by long dependency distance.

However, there is a different view that, instead of time-decay, it is interference that triggers forgetting [126,127]. Therefore, the processing difficulty caused by dependency distance can be interpreted as the result of the interference from the intervening words [128–130], for activation may be attenuated by similarity [123] through the Fan Effect [131], or, the cue word may raise, through spreading activation, the activation level of multiple words, leading to interference in retrieving [5,113]. For example, "*the student who love their books are sad*" may pose less processing difficulty than "*the student who love their mothers are sad*", for in the latter an animate noun "mother" may interfere with the subject "students". Thus, a linguistic regularity may be expected that long dependencies tend to involve interveners dissimilar to both the governing and the dependent word.

This possible regularity may account for the alleged contribution of case marking to reducing the processing difficulty of long dependencies [102], which seems to be supported by the psychological findings that long distance dependencies in German, a case-rich language, do not necessarily retard language processing [132]. One possible reason is that in SOV languages, the intervening object may lead to interference, which can be mended by marking the relations explicitly with different cases to rule out possible interference [99]. In brief, there could theoretically be interaction between degree of interference and dependency distance. In other words, a pattern may exist in languages that long dependency distance may tend not to involve intervening words similar to both the dependent word and the governing word, which is yet to be verified in empirical studies.

The decay-based account and the interference-based account are both built on the network organization of linguistic knowledge in mind [133] and the mechanism of spreading activation [131]. In fact, experimental studies have yielded results that seem to support both accounts and there is evidence that the effect of decay may well be independent of that of interference [41,45]. Theoretically, the decay-based accounts and the interference-based accounts could be integrated into unified models to capture the syntactic patterns concerning dependency distance [123].

In brief, though the universal pressure for DDM has molded an overwhelming majority of short dependencies, there are inevitably a small minority of long ones, motivated by other pressures or constraints. However, as a self-adapting synergetic system, a language can develop various patterns to cope with the small number of long-distance dependencies that are unavoidable in natural languages [99,111]. In other words, many long-distance dependencies that are actually used in natural languages are probably not difficult to process.

Interestingly, it has been proposed that high predictability or low surprisal may reduce processing difficulty [134], which seems to imply that the principle of least effort may lead to preference for long dependency distance, or dependency distance maximization, because expectation (predictability) may be enhanced or sharpened by longer distance, which is termed as anti-locality effect [123]. Some psychological experiments do show that in some cases longer dependency may sharpen expectation to facilitate processing [135–139]. Thus it is possible that the anti-locality effect has influence on human languages, or interacts with the locality effect to shape syntactic patterns. For example, a study of Hindi indicates that high expectation of verbs can cancel the locality effect, but "when expectation is weak, the facilitation effect disappears and a tendency towards a locality effect is seen" [50]. In Hindi, the adverbials intervening a complex predicate in fact facilitate processing since the sentence-final light verb of the complex predicate is highly predictable from the nominal predicate, which, also part of the complex predicate, precedes the intervening adverbials and the final light verb; this facilitation, however, disappears and the locality effect shows up in the cases of simple predicates where sentence-final predicate verb is not that predictable [50]. Another study has found a moderate correlation between the locality effect and the surprisal: dependency length makes good predications when surprisal fails, and vice versa [140].

Nevertheless, it should be pointed out that studies concerning the anti-locality effects are largely psychological experiments which are confined to very limited range of experimental subjects and materials. In contrast, big-data analysis of corpus seems to consistently report power-law distributions of dependency distance and short overall dependency distance, indicating a universal tendency toward dependency distance minimization. These consistent facts seem to suggest that the pressure for DDM probably plays a much more important role in shaping linguistic regularities than the pressure for DD maximization, i.e., predictability maximization. As to the small number of long-distance dependencies in natural languages, however, it might be expected that high predictability is another factor that may somewhat reduce the processing difficulty invoked by long distance.

DDM may be a very influential pressure in most modern languages whose sentences have evolved into increasing complexity and length, which account for the relatively short MDD and hence the majority of short dependencies in natural languages. At the same time, the findings that no language minimizes its MDD to the theoretical minimum and that there are always a small number of long-distance dependencies suggest the effect of other forces, like predictability maximization, pragmatic function, discourse coherence, iconicity, reliability, etc. For example, discourse coherence may often account for the occasional placements of adverbial between a subject and its predicate, like "*They, in order to succeed, work quite hard*", which adopts in fact a word order maximizing dependency distance. These forces and pressures may not have as potent effect on syntax and grammar as the pressure for DDM, but still have their roles to play.

In other words, the principle of least effort may get embodied in different requirements, with minimizing memory burden being a very important one. Similarly, minimizing memory burden may be realized by various syntactic patterns, with DDM being probably the dominant one, and even DDM can be realized in many different ways, with word ordering as the most conspicuous one. Given these facts and the resultant numerous possibilities, it is not surprising to find such diversity and variation among and within languages: slight contingent differences may drive language eventually into drastic differences and rich diversity. It is the interaction and the competition of these universal constraints and pressures that drive human languages to evolve, as complex systems, into today's rich diversity.

6. Concluding remarks

Human languages seem to present a preference for short dependency distance, which may be explained in terms of general cognitive constraint of limited working memory. Since this constraint is common for all human beings, it can be expected that this preference should be universal in various languages, that is, a linguistic universal driven by human beings. Big data analyses of various corpora have supported this assumption, claiming that overall dependency distance is significantly shorter than that of random languages and the dependency distance in natural languages all abides by similar distributions, with a majority of short dependencies and a handful of long ones. Human languages, as complex systems, have come up with various syntactic patterns to restrain dependency distance during an evolution into increasing sentential complexity and length.

However, the constraint of memory is probably one of constraints languages are subject to. Languages are largely meant for reliable, effective and efficient transmission of information, which may get embodied in different constraints. Sometimes, DDM has to give way to other needs, which may lead to the inevitability of the small number of long dependencies in various languages. Languages may adapt themselves to these long dependencies, thus developing some patterns to reduce the memory burden. Therefore, the sporadic long dependencies actually used are not necessarily difficult to process.

As a complex system, language is constantly adapting itself to several deep-level constraints that compete and interact with one another. These constraints shared by humanity give rise to the rich diversity of human languages, which are actually embodiments of hidden language universal.

Acknowledgements

This work was partly supported by the Social Science Foundation of Education Ministry of China (13YJC740112), the National Social Science Foundation of China (Grant No. 11&ZD188), the Fundamental Research Funds for the Central Universities (Program of Big Data PLUS Language Universals and Cognition, Zhejiang University), the MOE Project of the Center for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, and the Doctoral Funds of Anhui Jianzhu University.

References

[1] Saussure F. Course in general linguistics. New York: Philosophical Library; 1959.
[2] Tesnière L. Eléments de la syntaxe structurale. Paris: Klincksieck; 1959.
[3] Mel'čuk I. Levels of dependency in linguistic description: concepts and problems. In: Agel V, Eichinnger L, Eroms HW, Hellwig P, Herringer HJ, Lobin H, editors. Dependency and valency. An international handbook of contemporary research, vol. 1. Berlin–New York: De Gruyter; 2003. p. 189–229.
[4] Liu HT. Dependency grammar: from theory to practice. Beijing: Science Press; 2009 [in Chinese].
[5] Hudson RA. An introduction to word grammar. Cambridge: Cambridge University Press; 2010.
[6] Liu HT. Dependency distance as a metric of language comprehension difficulty. J Cogn Sci 2008;9(2):159–91.
[7] Futrell R, Mahowald K, Gibson E. Large-scale evidence for dependency length minimization in 37 languages. Proc Natl Acad Sci USA 2015;112(33):10336–41.
[8] Zipf G. Human behavior and the principle of least effort: an introduction to human ecology. New York: Hafner; 1949.
[9] Liu HT. Probability distribution of dependency distance. Glottometrics 2007;15:1–12.
[10] Ferrer-i-Cancho R. Euclidean distance between syntactically linked words. Phys Rev A 2004;70:056135.
[11] Temperley D. Dependency length minimization in natural and artificial languages. J Quant Linguist 2008;15:256–82.
[12] Temperley D. Minimization of dependency length in written English. Cognition 2007;105:300–33.
[13] Gildea D, Temperley D. Do grammars minimize dependency length? Cogn Sci 2010;34:286–310.
[14] Bhhaghel O. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. Indogermanische Forschungen; 1909. p. 25.110–42.
[15] Yngve V. A model and an hypothesis for language structure. Proc Am Philos Soc 1960;104:444–66.
[16] Frazier L, Rayner K. Resolution of syntactic category ambiguities: eye movements in parsing lexically ambiguous sentences. J Mem Lang 1987;26:505–26.
[17] Gibson E. A computational theory of human linguistic processing: memory limitations and processing break down [doctoral dissertation], Pittsburgh, PA: Carnegie Mellon University; 1991.
[18] Heringer HJ, Strecker B, Wimmer R. Syntax Fragen-Lösungen Alternativen. München: Wilhelm Fink Verlag; 1980.
[19] Hudson R. Measuring syntactic difficulty. Available from: http://dickhudson.com/wp-content/uploads/2013/07/Difficulty.pdf [2016-9-6].
[20] Brown J. Some tests of the decay theory of immediate memory. J Exp Psychol 1958;10:173–89.
[21] Baddeley AD, Hitch GJ. Working memory. In: Bower G, editor. Recent advances in learning and motivation. Academic Press; 1974. p. 47–89.
[22] Hawkins JA. A performance theory of order and constituency. Cambridge: Cambridge University Press; 1994.
[23] Hawkins JA. Efficiency and complexity in grammars. Oxford: Oxford University Press; 2004.
[24] Gibson E. Linguistic complexity: locality of syntactic dependencies. Cognition 1998;68:1–76.
[25] Gibson E. The dependency locality theory: a distance-based theory of linguistic complexity. In: Marantz A, Miyashita Y, O'Neil W, editors. Image, language, brain. Boston: MIT Press; 2000. p. 95–126.
[26] Robins RH. A short history of linguistics. London and New York: Longman; 1967.
[27] Arnauld A, Lancelot C. General and rational grammar: the Port-Royal grammar. The Hague: Mouton; 1975.
[28] Jespersen O. The philosophy of grammar. London: Allen and Unwin; 1924.
[29] Chomsky N. New horizons in the studies of language and mind. Cambridge: Cambridge University Press; 2000.
[30] Greenberg JH. Universals of language. Cambridge, Mass: MIT Press; 1963.
[31] Croft W. Typology and universals. Cambridge: Cambridge University Press; 2002.
[32] Ferrer-i-Cancho R. Beyond description. Comment on "Approaching human language with complex networks" by Cong and Liu. Phys Life Rev 2014;11(4):621–3.
[33] Haken H. Synergetics, an introduction: nonequilibrium phase transitions and self-organization in physics, chemistry, and biology. New York: Springer-Verlag; 1983.
[34] Haken H. Advanced synergetics: instability hierarchies of self-organizing systems and devices. New York: Springer-Verlag; 1993.
[35] Kauffman S. Origins of order: self organisation and selection in evolution. Oxford: Oxford University Press; 1993.
[36] Liu HT. Language is more a human-driven system than a semiotic system. Comment on "Modeling language evolution: examples and predictions". Phys Life Rev 2014;11:309–10.
[37] Hsiao F, Gibson E. Processing relative clauses in Chinese. Cognition 2003;90:3–27.
[38] Grodner D, Gibson E. Some consequences of the serial nature of linguistic input. Cogn Sci 2005;29(2):261–90.
[39] Levy R, Fedorenko E, Gibson E. The syntactic complexity of Russian relative clauses. J Mem Lang 2013;69:461–95.
[40] Gibson E, Processing Wu I. Chinese relative clauses in context. Lang Cogn Neurosci 2013;28:125–55.
[41] Fedorenko E, Piantadosi S, Gibson E. Processing relative clauses in supportive contexts. Cognitive Sci 2012;36:471–97.
[42] Chen B, Ning A, Bi H, Dunlap S. Chinese subject-relative clauses are more difficult to process than the object-relative clauses. Acta Psychol 2008;129:616.
[43] Lin Y, Garnsey SM. Animacy and the resolution of temporary ambiguity in relative clause comprehension in Mandarin. In: Yamashita H, Hirose Y, Packard J, editors. Processing and producing head-final structures. Dordrecht, The Netherlands: Springer; 2010. p. 241–75.
[44] Levy R, Keller F. Expectation and locality effects in German verb-final structures. J Mem Lang 2013;68(2):199–222.
[45] Bartek B, Lewis RL, Vasishth S, Smith MR. In search of on-line locality effects in sentence comprehension. J Exp Psychol Learn 2011;37(5):1178–98.
[46] Fedorenko E, Woodbury R, Gibson E. Direct evidence of memory retrieval as a source of difficulty in non-local dependencies in language. Cogn Sci 2013;37(2):378–94.
[47] Levy R, Fedorenko E, Breen M, Gibson E. The processing of extraposed structures in English. Cognition 2012;122:12–36.

[48] Hofmeister P, Jaeger TF, Sag IA, Arnon I, Snider N. Locality and accessibility in Wh-questions. In: Featherston S, Sternefeld W, editors. Roots: linguistics in search of its evidential base. Berlin: Mouton de Gruyter; 2007.
[49] Hofmeister P, Sag IA. Cognitive constraints and Island effects. Language 2010;86:366–415.
[50] Husain HS, Vasishth S, Srinivasan N. Strong expectations cancel locality effects: evidence from Hindi. PLoS ONE 2014;9(7):e100986.
[51] Lin YW. Locality versus anti-locality effects in Mandarin sentence comprehension. In: Zhuo Jing-Schmidt, editor. Proceedings of the 23rd North American conference on Chinese linguistics (NACCL-23), vol. 1. Eugene: University of Oregon; 2011. p. 200–14.
[52] Misyak JB, Christiansen MH. When 'more' in statistical learning means 'less' in language: individual differences in predictive processing of adjacent dependencies. In: Catrambone R, Ohlsson S, editors. Proceedings of the 32nd annual cognitive science society conference. Austin, TX: Cognitive Science Society; 2010. p. 2686–91.
[53] Nicol JL, Forster KI, Veres C. Subject–verb agreement processes in comprehension. J Mem Lang 1997;36:569–87.
[54] Pearlmutter NJ, Garnsey SM, Bock K. Agreement processes in sentence comprehension. J Mem Lang 1999;41:427–56.
[55] Vasishth S, Chen Z, Li Q, Processing Guo G. Chinese relative clauses: evidence for the subject-relative advantage. PLoS ONE 2013;8(10):e77006. http://dx.doi.org/10.1371/journal.pone.0077006.
[56] Lin CC. Relative-clause processing in typologically distinct languages: a universal parsing account [doctoral dissertation], Tucson, AZ, USA: University of Arizona; 2006.
[57] Lin CC, Bever TG. Subject preference in the processing of relative clauses in Chinese. In: Baumer D, Montero D, Scanlon M, editors. Proceedings of the 25th west coast conference on formal linguistics. Somerville, MA: Cascadilla Proceedings Project; 2006. p. 254–60.
[58] Mayer-Schonberger V, Cukier K. BIG DATA: a revolution that will transform how we live, work, and think. New York: John Murray General Publishing Division; 2013.
[59] Hiranuma S. Syntactic difficulty in English and Japanese: a textual study. UCL Working Pap Linguist 1999;11:309–22.
[60] Eppler E. Dependency distance and bilingual language use: evidence from German/English and Chinese/English data. In: Hajičová Eva, Gerdes Kim, Wanner Leo, editors. Proceedings of the second international conference on dependency linguistics (DepLing2013). Prague: Association for Computational Linguistics; 2013. p. 78–87.
[61] Lu Q, Liu HT. Does dependency distance distribute regularly? J Zhejiang Univ (Humanit Soc Sci) 2015;4:63–76. http://dx.doi.org/10.3785/j.issn.1008-942X.CN33-6000/C.2015.12.231 [in Chinese].
[62] Ferrer-i-Cancho R, Solè RV. Least effort and the origins of scaling in human language. Proc Natl Acad Sci 2003;100(3):788–91.
[63] Jiang JY, Liu HT. The effects of sentence length on dependency distance, dependency direction and the implications. Lang Sci 2015;50:93–104.
[64] Altmann-Fitter. Altmann-Fitter user guide. The third version. Available from: http://www.ram-verlag.eu/wp-content/uploads/2013/10/Fitter-User-Guide.pdf, 2013 [2016-10-12].
[65] Wimmer G, Altmann G. Thesaurus of univariate discrete probability distributions. Essen: STAMM Verlag; 1999.
[66] Wang YQ, Liu HT. The effects of genre on dependency distance and dependency direction. Lang Sci 2017;59:135–47.
[67] Ferrer-i-Cancho R, Liu HT. The risks of mixing dependency lengths from sequences of different length. Glottotheory 2014;5(2):143–55.
[68] Ferrer-i-Cancho R. Why do syntactic links not cross? Europhys Lett 2006;76:1228–35.
[69] Ferrer-i-Cancho R. Some word order biases from limited brain resources. A mathematical approach. Adv Complex Syst 2008;11(3):394–414.
[70] Park YA, Levy R. Minimal-length linearizations for mildly context-sensitive dependency trees. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the association for computational linguistics. Boulder: The Association for Computational Linguistics; 2009. p. 335–43.
[71] Crabbé B, Gulordava K, Merlo P. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In: ACL proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (short papers). Beijing: The Association for Computational Linguistics; 2015. p. 477–82.
[72] Wang H, Liu HT. The effect of length and complexity on constituent ordering in written English. Poznań Stud Contemp Linguist 2014;50(4):477–94.
[73] Quirk R, Greenbaum S, Leech G, Svartvik J. A grammar of contemporary English. London: Longman; 1972.
[74] Bresnan J, Cueni A, Nikitina T, Baayen H. Predicting the dative alternation. In: Boume G, Kraemer I, Zwarts J, editors. Cognitive foundations of interpretation. Amsterdam: Royal Netherlands Academy of Science; 2007. p. 69–94.
[75] Wasow T. End-weight from the speaker's perspective. J Psycholinguist Res 1997;26:347–61.
[76] Hoffmann C. Word order and the principle of "Early Immediate Constituents" (EIC). J Quant Linguist 1999;6(2):108–16.
[77] Stallings L, MacDonald MC, O'Seaghdha PG. Phrasal ordering constraints in sentence production: phrase length and verb disposition in heavy-NP shift. J Mem Lang 1998;39:392–417.
[78] Arnold J, Wasow T, Losongco A, Grinstrom R. Heaviness vs. newness: the effects of structural complexity and discourse status on constituent ordering. Language 2000;76:28–55.
[79] Yamashita H, Chang F. Long before short preference in the production of a head-final language. Cognition 2001;81:B45–55.
[80] Faghiri P, Samvelian P. Constituent ordering in Persian and the weight factor. In: Christopher P, editor. Empirical issues in syntax and semantics 10 (EISS10); 2014. Available from: http://www.cssp.cnrs.fr/eiss10/eiss10_faghiri-and-samvelian.pdf [2016-9-15].
[81] Choi HW. Length and order: a corpus study of Korean dative-accusative construction. Discourse Cogn 2007;14(3):207–27. Available from: http://home.ewha.ac.kr/club/hwchoi/public_html/Content/Choi_DisCog07.pdf [2016-3-20].
[82] Choi HW. Optimizing structure in context: scrambling and information structure. Stanford, CA: CSLI Publications; 1999.
[83] Köhler R. Synergetic linguistics. In: Köhler R, Altmann G, Piotrowski RG, editors. Quantitative linguistics: an international handbook. Berlin: Mouton de Gruyter; 2005. p. 760–74.
[84] Ueno M, Polinsky M. Does headedness affect processing? A new look at the vo-ov contrast. J Linguist 2009;45:675–710.

[85] Ferrer-i-Cancho R. Hubiness, length and crossings and their relationships in dependency trees. Glottometrics 2013;25:1–21.
[86] Ferrer-i-Cancho R, Gómez-Rodríguez C. Crossings as a side effect of dependency lengths. Complexity 2016;21(S2):320–8.
[87] Ferrer-i-Cancho R. A stronger null hypothesis for crossing dependencies. Europhys Lett 2014;108:58003.
[88] Ferrer-i-Cancho R, Gómez-Rodríguez C. The scarcity of crossing dependencies: a direct outcome of a specific constraint?. arXiv:1601.03210 [cs.CL], 2016. Available from: https://arxiv.org/pdf/1601.03210v1 [2015-9-10].
[89] Lu Q, Liu HT. A quantitative study on the relationship between crossing and distance of human language. J Shanxi Univ (Philos Soc Sci Edition) 2016;39(4):49–56 [in Chinese].
[90] Esteban JL, Ferrer-i-Cancho R, Gómez-Rodríguez C. The scaling of the minimum sum of edge lengths in uniformly random trees. J Stat Mech 2016;2016(6). http://dx.doi.org/10.1088/1742-5468/2016/06/063401.
[91] Ferrer-i-Cancho R. Random crossings in dependency trees. arXiv:1305.4561 [cs.CL], 2013. Available from: https://arxiv.org/ftp/arxiv/papers/1305/1305.4561.pdf [2016-8-20].
[92] Lu Q, Xu CS, Liu HT. Can Chunking reduce syntactic complexity of natural languages?. Complexity 2016;21(S2):33–41.
[93] Oya M. Degree centralities, closeness centralities, and dependency distances of different genres of texts. In: Selected papers from the 17th conference of pan-pacific applied linguistics; 2013. p. 42–53. Available from: http://www.paaljapan.org/conference2012/pdf/006oya.pdf [2016-4-20].
[94] Jing YQ, Liu HT. Mean hierarchical distance: augmenting mean dependency distance. In: Proceedings of the third international conference on dependency linguistics (DepLing2015). Uppsala: Association for Computational Linguistics; 2015. p. 161–70.
[95] Köhler R. Quantitative syntax analysis. Berlin/Boston: Mouton De Gruyter; 2012.
[96] Gell-Mann M, Ruhlen M. The origin and evolution of word order. Proc Natl Acad Sci USA 2011;108(42):17290–5.
[97] Reali F, Christiansen MH. Sequential learning and the interaction between biological and linguistic adaptation in language evolution. Interact Stud 2009;10:5–30.
[98] Langus A, Nespor M. Cognitive systems struggling for word order. Cogn Psychol 2010;60(4):291–318.
[99] Gibson E, Piantadosi S, Brink K, Bergen L, Lim E, Saxe R. A noisy-channel account of cross-linguistic word order variation. Psychol Sci 2013;4(7):1079–88.
[100] Ferrer-i-Cancho R. Why might SOV be initially preferred and then lost or recovered? A theoretical framework. In: Cartmill EA, Roberts S, Lyn H, Cornish H, editors. The evolution of language – proceedings of the 10th international conference (EVOLANG10). Singapore: World Scientific; 2014. p. 66–73.
[101] Givon T. The genesis of syntactic complexity. Amsterdam–Philadelphia: Benjamins; 2009.
[102] Tily H. The role of processing complexity in word order variation and change [doctoral dissertation], Stanford, California: Stanford University; 2010.
[103] Ferrer-i-Cancho R. The placement of the head that minimizes online memory: a complex systems approach. Lang Dyn Chang 2013;5(1):114–37.
[104] Liu HT, Xu CS. Quantitative typological analysis of Romance languages. Poznań Stud Contemp Linguist 2012;48(4):597–625.
[105] Lehmann WP. A structural principle of language and its implications. Language 1973;49(1):47–66.
[106] White M, Rajkumar R. Minimal dependency length in realization ranking. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island: Association for Computational Linguistics; 2012. p. 244–55.
[107] Eisner J, Smith NA. Favor short dependencies: parsing with soft and hard constraints on dependency length. In: Bunt H, Merlo P, Nivre J, editors. Trends in parsing technology: dependency parsing, domain adaptation, and deep parsing. Speech and language technology, vol. 43. Springer; 2011. p. 121–50.
[108] Boston MF. A computational model of cognitive constraints in syntactic locality [doctoral dissertation], Ithaca, New York: Cornell University; 2012.
[109] Christiansen MH, Chater N. The now-or-never bottleneck: a fundamental constraint on language. Behav Brain Sci 2016;39. http://dx.doi.org/10.1017/S0140525X1500031X.
[110] Liu BL. Diachronic shifts of Chinese: a dependency treebank-based study [doctoral dissertation], Beijing: Communication University of China; 2013 [in Chinese].
[111] Xu CS, Liu HT. Can familiarity lessen the effect of locality? A case study of Mandarin Chinese subjects and the following adverbials. Poznań Stud Contem Linguist 2015;51(3):463–86.
[112] Miller GA, Chomsky N. Finitary models of language users. In: Luce D, Bush R, Galanter E, editors. Handbook of mathematical psychology Volume II. John Wiley & Sons; 1963. p. 419–91.
[113] Hudson RA. Language networks: a new word grammar. Cambridge: Cambridge University Press; 2007.
[114] Wang L, Liu HT. Syntactic variations in Chinese–English code-switching. Lingua 2013;123:58–73.
[115] Reali F, Christiansen MH. Word-chunk frequencies affect the processing of pronominal object-relative clauses. Q J Exp Psychol 2007;60:161–70.
[116] Xu CS. The use and the omission of Chinese conjunction "er". J Shanxi Univ (Philos Soc Sci Edition) 2015;38(2):55–61 [in Chinese].
[117] Jaeger TF. Redundancy and reduction: speakers manage syntactic information density. Cogn Psychol 2010;61:23–62.
[118] Peña M, Bonatti LL, Nespor M, Mehler J. Signal-driven computations in speech processing. Science 2002;298:604–7.
[119] Endress AD, Dehaene-Lambertz G, Mehler J. Perceptual constraints and the learnability of simple grammars. Cognition 2007;105(3):577–614.
[120] Endress AD, Mehler J. Primitive computations in speech processing. Q J Exp Psychol 2009;62(11):2187–209.
[121] Sousa DA. How the brain learns. 4th ed. Thousand Oaks: Corwin Press; 2011.
[122] Lewis RL, Vasishth S. An activation-based model of sentence processing as skilled memory retrieval. Cogn Sci 2005;29:1–45.

[123] Vasishth S, Lewis RL. Argument-head distance and processing complexity: explaining both locality and anti-locality effects. Language 2006;82(4):767–94.
[124] Xu CS. Chinese long distance dependencies [doctoral dissertation], Beijing: Communication University of China; 2013 [in Chinese].
[125] Liu HT, Feng ZW. Probabilistic valency pattern theory for natural language processing. Language Sci 2007;6(3):32–41 [in Chinese].
[126] White KG. Dissociation of short-term forgetting from the passage of time. J Exp Psychol Learn 2012;38:255–9.
[127] Oberauer K, Lewandowsky S. Evidence against decay in verbal working memory. J Exp Psychol Gen 2013;142(2):380–411. http://dx.doi.org/10.1037/a0029588.
[128] VanDyke JA. Interference effects from grammatically unavailable constituents during sentence processing. J Exp Psychol Learn 2007;33(2):407–30.
[129] VanDyke JA, Lewis RL. Distinguishing effects of structure and decay on attachment and repair: a retrieval interference theory of recovery from misanalyzed ambiguities. J Mem Lang 2003;49(3):285–316.
[130] Levy R. Memory and surprisal in human sentence comprehension. In: van Gompel R, Roger PG, editors. Sentence processing. Hove: Psychology Press; 2013. p. 78–114.
[131] Anderson JR, Reder LM. The fan effect: new results and new theories. J Exp Psychol Gen 1999;128:186–97.
[132] Konieczny L. Locality and parsing complexity. J Psycholinguist Res 2000;29(6):628–45.
[133] Liu HT, Cong J. Empirical characterization of modern Chinese as a multi-level system from the complex network approach. J Chin Linguist 2014;42(1):1–38.
[134] Levy R. Expectation-based syntactic comprehension. Cognition 2008;106:1126–77.
[135] Nakatani K, Gibson E. Distinguishing theories of syntactic expectation cost in sentence comprehension: evidence from Japanese. Linguistics 2008;46(1):63–87.
[136] Jäger L, Chen Z, Li Q, Lin CC, Vasishth S. The subject-relative advantage in Chinese: evidence for expectation-based processing. J Mem Lang 2015;79–80:97–120.
[137] Vasishth S, Chen Z, Li Q, Guo G. Processing Chinese relative clauses: evidence for the subject-relative advantage. PLoS ONE 2013;8(10):e77006. http://dx.doi.org/10.1371/journal.pone.0077006.
[138] Vasishth S, Drenhaus H. Locality in German. Dialogue Discourse 2011;1:59–82. http://dx.doi.org/10.5087/dad.2011.104.
[139] Nicenboim B, Vasishth S, Gattei C, Sigman M, Kliegl R. Working memory differences in long-distance dependency resolution. Front Psychol 2015;6:312. http://dx.doi.org/10.3389/fpsyg.2015.00312.
[140] Rajkumara R, van Schijndel M, White M, Schuler W. Investigating locality effects and surprisal in written English syntactic choice phenomena. Cognition 2016;155:204–32.