Non-crossing Dependencies: Least Effort, not Grammar

Ramon Ferrer-i-Cancho

Abstract The use of null hypotheses (in a statistical sense) is common in hard sciences but not in theoretical linguistics. Here the null hypothesis that the low frequency of syntactic dependency crossings is expected by an arbitrary ordering of words is rejected. It is shown that this would require star dependency structures, which are both unrealistic and too restrictive. The hypothesis of the limited resources of the human brain is revisited. Stronger null hypotheses taking into account actual dependency lengths for the likelihood of crossings are presented. Those hypotheses suggests that crossings are likely to reduce when dependencies are shortened. A hypothesis based on pressure to reduce dependency lengths is more parsimonious than a principle of minimization of crossings or a grammatical ban that is totally dissociated from the general and non-linguistic principle of economy.

1 Introduction

A substantial subset of theoretical frameworks under the general umbrella of ”generative grammar” or ”generative linguistics” have been kidnapped by the idea that a deep theory of syntax requires that one neglects the statistical properties of the system [2] and abstracts away from functional factors such as the limited resources of the brain [3].

This radical assumption disguised as intelligent abstraction led to the distinction between competence and performance (see [4, Sect. 2.4] for a historical perspective from generative grammar), a dichotomy that is sometimes regarded as a soft methodological division [4, pp. 34] or as theoretically unmotivated [5]. A sister

Ramon Ferrer-i-Cancho
Complexity and Quantitative Linguistics Lab, LARCA Research Group. Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC). Campus Nord, Edifici Omega, Jordi Girona Salgado 1-3. 08034 Barcelona, Catalonia (Spain), e-mail: rferrericancho@cs.upc.edu.

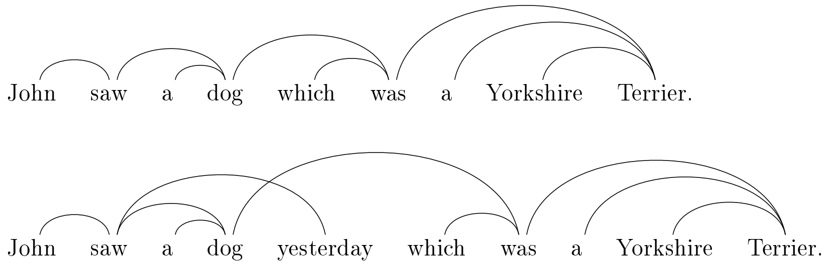

John saw a dog which was a Yorkshire Terrier.
John saw a dog yesterday which was a Yorkshire Terrier.

Fig. 1 **Top** an English sentence without crossings. **Bottom** a variant of the previous sentence with one dependency crossing (the dependency between ”saw” and ”yesterday” crosses the dependency between ”dog” and ”was” and vice versa). Adapted from [1].

radical dichotomy is the division between grammar and usage [6]. A revision of those views has led to the proposal of competence-plus, *”a package consisting of the familiar recursive statements of grammatical rules, plus a set of numerical constraints on the products of those rules”* [7]. Interestingly, certain approaches reconcile competence with performance by regarding grammar as a store of ”frozen” or ”fixed” performance preferences [8, p. 3] or by opening the set of numerical constraints of competence-plus to performance factors [7]. Other examples of approaches that reject the dichotomous view of language are emergent grammar [9], synergetic linguistics [10] or probabilistic syntax [11]. The challenges of the competence/performance are not specific to generative linguistics. For instance, *”the competence/performance distinction is also embodied in many symbolic models of language processing”* [12] and integrated with some refinements in language evolution research [7].

Again from the perspective of standard model selection [13], the competence/-performance dichotomy, even in soft versions, has a serious risk: if a more parsimonious theory exists based on performance, one that has the same or even superior explanatory power, it may not be discovered and if so, it will not be sufficiently endorsed. Astonishingly, linguistic theories that belittle the role of the limited resources of the human brain for structural constraints of syntax are presented as minimalistic (e.g., [14]). In contrast, standard model selection favors theories with a good compromise between simplicity (often coming from a suitable abstraction or idealization) and explanatory power [13].

A follower of the competence-performance split may consider that the opponents are unable to think in sufficiently abstract terms: opponents are being side-tracked by actual language use and limited computational resources and do not focus on the essence of syntax (in those views the essence of syntax is grammar [6] or certain features of such as recursion [15]; in other approaches the essence of syntax is not grammar but dependencies [16, 17]). The proponents of the split doctrine have not hesitated either to advertise functional approaches to linguistic theory as wrong [18] or to attempt to dismantle attempts to turn research on language or communication more quantitative (e.g., [2, 19, 20]). A real scientist however, will ask for the quality of a theory or a hypothesis in terms of the accuracy of its definitions, its testability, the statistical analyzes that have been performed to support it, the null hypotheses, the trade-off between explanatory power and parsimony of the theory, and so on.

If the limited resources of the brain are denied, one might be forced to blame grammar for the occurrence of certain patterns. Using standard model selection terms [13], forwarding the responsibility to grammar implies the addition of more parameters to the model, indeed unnecessary parameters, as it will be shown here through a concrete phenomenon. The focus of the current article is a striking pattern of syntactic dependency trees of sentences that was reported in the 1960s: dependencies between words normally do not cross when drawn over the sentence [21, 22] (e.g., Fig. 1). The problem of dependency crossings looks purely linguistic but it goes beyond human language: crossings have also been investigated in dependency networks where vertices are occurrences of nucleotides A, G, U, and C and edges are U-G and Watson-Crick base pairs, i.e. A-U, G-C [23]. Having in mind various domains of application helps a researcher to apply the right level of abstraction. Becoming a specialist in human language or certain linguistic phenomena helps to find locally optimal theories, causes the illusion of scientific success when becoming the world expert of a certain topic but does not necessarily produce compact, coherent, general and elegant theories.

Here new light is shed on the origins of non-crossing dependencies by means of two fundamental tools of the scientific method: null hypotheses and unrestricted parsimony (unrestricted parsimony in the sense of being a priori open to favor theories that make fewer assumptions; not in the sense that parsimony has to be favored neglecting explanatory power). Unfortunately, the definition of null hypotheses (in a statistical sense) is rare in theoretical linguistics (although it is fundamental in biology or medicine). Even in the context of quantitative linguistics research, clearly defined null hypotheses or baselines are present in certain investigations, e.g., [24, 25] but are missing in others e.g., [26, 27]. When present, they are not always tested rigorously [28]. In the context of quantitative research, claims about the efficiency of language have been made lacking a measure of cost and evidence that such a cost is below chance [29]. A deep theory of language requires (at least) metrics of efficiency, tests of their significance and an understanding of the relationship between the minimization of the costs that they define and the emergence of the target patterns, e.g., Zipf's law of abbreviation [24].

To our knowledge, claims for the existence of a universal grammar have never been defended by means of a null hypothesis (in a statistical sense), e.g., [4, 30], and a baseline is missing in research where grammar is seen as a conventionalization of performance constraints [8] or in research where competence is complemented with quantitative constraints [7]. As for the latter, baselines would help one to determine which of those constraints must be stored by grammar or competence-plus.

The first question that a syntactician should ask as a scientist when investigating the origins of a syntactic property X is: could X happen by chance? The question is equivalent to asking if grammar (in the sense of some extra knowledge) or specific genes are needed to explain property X. Accordingly, the major question that this article aims to answer is: could the low frequency of crossings in syntactic dependency trees be explained by chance, maybe involving general computational constraints of the human brain?

The remainder of the article is organized as follows. Sect. 2 reviews our minimalistic approach to the syntactic dependency structure of sentences. Sect. 3 considers the null hypothesis of a random ordering of the words of the sentence and shows that keeping the expected number of crossings small requires unrealistic constraints on the ensemble of possible dependency trees (only star trees would be possible). Sect. 4 considers alternative hypotheses, discarding the vague or heavy hypothesis of grammar and focusing on two major hypotheses: a principle of minimization of crossings and a principle of minimization of the sum of dependency lengths. The analysis suggests that the number of crossing and the sum of dependency lengths are not perfectly correlated but their correlation is strong. Of the two principles, dependency length minimization offers a more parsimonious account of many more linguistic phenomena. Interestingly, that principle is motivated by the need of minimizing cognitive effort. A challenge for the hypothesis that the rather small number of crossings of real sentence is a side-effect of minimization of dependency lengths is (a) determining the degree of that minimization that the real number of crossings requires and (b) if that degree is realistic. Sect. 5 presents a stronger null hypothesis that addresses the challenge with knowledge about edge lengths. That null hypothesis allows one to predict the number of crossings when the length of one of the edges potentially involved in a crossing is known but words are arranged at random. Thus, that predictor uses the actual dependency lengths to estimate the number of crossings. Interestingly, that predictor provides further support for a strong correlation between crossings and dependency lengths: analytical arguments suggest that it is likely that a reduction of dependency lengths causes a drop in the number of crossings. Sect. 6 considers another predictor based on a stronger null hypothesis where the sum of dependency lengths is given but words are arranged at random. Preliminary numerical results indicate a strong correlation between the mean number of crossings and the sum of dependency lengths over all the possible orderings of the words of real sentences. Interestingly, that null hypothesis leads to a predictor that requires less information about a real sentence than the previous predictor (only the sum of dependency lengths is needed) and paves the way to understanding the rather low number of crossings in real sentences as a consequence of global cognitive constraints on dependency lengths. Sect. 7 compares the predictions of the three predictors on a small set of sentences. The results suggest that the predictor based on the sum of dependency lengths is the best candidate. There it is also demonstrated that p-value testing can be used to investigate the adequacy of the best candidate. Interestingly, the best candidate was not rejected in that sample of sentences. Finally, Sect. 8 reviews and discusses the idea that least effort, not grammar, is the reason for the small number of crossings of real sentences.

2 The Syntactic Dependency Structure of Sentences

This article borrows the minimalistic approach to the syntactic dependency structure of sentences of dependency grammar [31, 32] and recent progress in cognitive sciences [17]:

- No hierarchical phrase structure is assumed in the sense that the structure of a sentence is defined simply as a tree where vertices are words and edges are syntactic dependencies. This is a fundamental assumption of our approach: tentatively, the network defining the dependencies between words might be a disconnected forest or a graph with cycles (these are possibilities that have not been sufficiently investigated) [31]. A general theory of crossings in nature cannot obviate the fact that RNA structures cannot be modeled with trees but can be modeled with forests [23] [1]. Although the choice of a tree of words as the reference model for sentence structure (e.g., [1]) is to some extent arbitrary, a tree is optimal for being the kind of network that is able to connect all words with the smallest amount of edges [34].
- Words establish direct relationships that are not necessarily mediated by syntactic categories (non-terminals in the phrase structure formalism and generative grammar evolutions). This skepticism about syntactic categories (as entities by its own, not epiphenomena) goes beyond dependency grammar, e.g., construction grammar [35].

Along the lines of [17], link direction is irrelevant for the arguments in this article. Even within the dependency grammar formalism, dependencies are believed to be directed (from heads to modifiers/complements) [32, 31]. A minimalistic approach to dependency syntax should not obviate the fact that the accuracy of dependency parsing improves if link direction is neglected [36].

3 The Null Hypothesis

Let n be the number of vertices of a tree. Let k_i be the degree of the i-th vertex of a tree and $k_1,...,k_i,...,k_n$ its degree. By K_α, we denote

$$K_\alpha = \sum_{i=1}^{n} k_i^\alpha, \tag{1}$$

where α is a natural number. In a tree, K_1 only depends on n, i.e. [37],

$$K_1 = 2(n-1) \tag{2}$$

[1] In those RNA structures, vertex degrees do not exceed one [23] and thus cycles are not possible but connectedness is not either (the handshaking lemma [33, p. 4] indicates that such a graph cannot have more than $n/2$ edges, being n the number of vertices, and thus cannot be connected because that needs at least $n-1$ edges).

and thus the 1st moment of degree is

$$\langle k \rangle = \frac{K_1}{n} = 2 - \frac{2}{n}. \tag{3}$$

Let $E_0[C]$ be the expected number of crossings in a random linear arrangement of a dependency tree with a given degree sequence, i.e. [38]

$$E_0[C] = \frac{n}{6}\left(n - 1 - \left\langle k^2 \right\rangle\right), \tag{4}$$

where $\left\langle k^2 \right\rangle$ is the 2nd moment of degree, i.e.

$$\left\langle k^2 \right\rangle = \frac{K_2}{n}. \tag{5}$$

Thus, the expected number of crossings depends on the number of vertices (n) and the 2nd moment of degree ($\left\langle k^2 \right\rangle$). The higher the hubiness (the higher $\left\langle k^2 \right\rangle$) the lower the expected number of crossings.

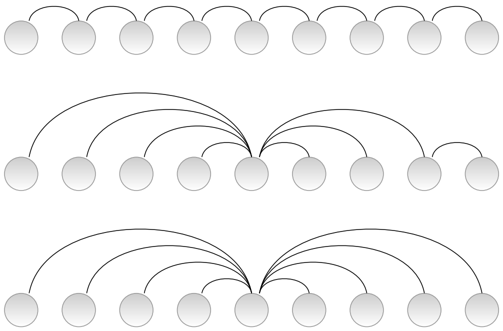


Fig. 2 Linear arrangements of trees of nine vertices. **Top** a linear tree. **Center** quasi-star tree. **Bottom** star tree.

A star tree is a tree with a vertex of degree $n-1$ while a linear tree is a tree where no vertex degree exceeds two [38] (Fig. 2). For a given number of vertices, $E_0[C]$ is minimized by star trees, for which $E_0[C] = 0$, whereas $E_0[C]$ is maximized by linear trees, for which [39, 38]

$$E_0[C] = \frac{1}{6}n(n-5) + 1. \tag{6}$$

As $E_0[C]$ depends on the $\left\langle k^2 \right\rangle$ of the tree, the null hypothesis that the tree structures are chosen uniformly at random among all possible labeled trees is considered next. The Aldous-Brother algorithm allows one to generate uniformly random labeled spanning trees from a graph [40, 41]. Here a complete graph is assumed to be the source for the spanning trees. A low number of crossings cannot be attributed to grammar if $E_0[C]$ is low.

$E_0[C]$ is the expectation of C given a degree sequence. Indeed, that expectation can be obtained just from knowledge about $\left\langle k^2 \right\rangle$ and n (Eq. 4). The expectation of C for uniformly random labeled trees is

Proposition 1.

$$\begin{aligned} E[E_0[C]] &= \frac{1}{6}(n-1)\left(n-5+\frac{6}{n}\right) \\ &= \frac{n^2}{6} - n + \frac{11}{6} - \frac{1}{n}. \end{aligned} \tag{7}$$

Proof. On the one hand, the degree variance for uniformly random labeled trees is [42, 37]

$$V[k] = \left\langle k^2 \right\rangle - \langle k \rangle^2 = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right). \tag{8}$$

Applying Eq. 3, it is obtained

$$\left\langle k^2 \right\rangle = \left(1-\frac{1}{n}\right)\left(5-\frac{6}{n}\right). \tag{9}$$

On the other hand,

$$\begin{aligned} E[E_0[C]] &= E\left[\frac{n}{6}\left(n-1-\left\langle k^2 \right\rangle\right)\right] && \text{applying Eq. 4} \\ &= \frac{n}{6}\left(n-1-E\left[\left\langle k^2 \right\rangle\right]\right) \\ &= \frac{n}{6}\left(n-1-\left(1-\frac{1}{n}\right)\left(5-\frac{6}{n}\right)\right) && \text{applying Eq. 9} \\ &= \frac{1}{6}(n-1)\left(n-5+\frac{6}{n}\right). \end{aligned}$$

□

Fig. 3 shows that uniformly random labeled trees exhibit a high value of $E_0[C]$ that is near the upper bound defined by linear trees. Thus, it is unlikely that the rather low frequency of crossings in real syntactic dependency trees [32, 43] is due to uniform sampling of the space of labeled trees. However, one cannot exclude the possibility that real dependency trees belong to a subclass of random trees for which $E_0[C]$ is low (e.g., the uniformly random trees may not be spanning trees of a complete graph). This possibility is explored next.

A quasi-star tree is defined as a tree with one vertex of degree $n-2$, one vertex of degree 2 and the remainder of vertices of degree 1 (Fig. 2). A quasi-star tree needs $n \geq 3$ to exist (Appendix). The sum of squared degrees of such a tree is (Appendix)

$$K_2^{\text{quasi}} = n^2 - 3n + 6 \tag{10}$$

and thus the degree 2nd moment of a quasi-star tree is

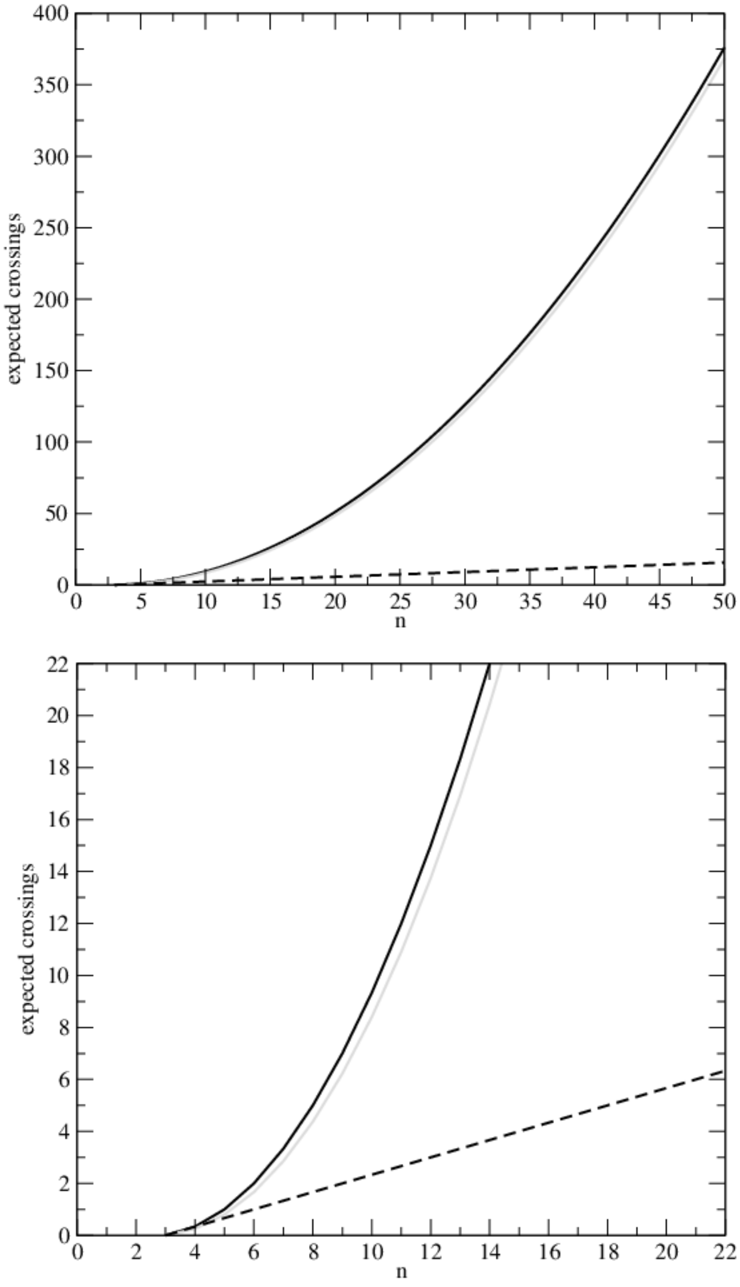


Fig. 3 The expected number of crossings as a function of the number of vertices of the tree (n) in random linear arrangements of vertices for linear trees (*black solid line*), uniformly random labeled trees (*gray line*) and quasi-star trees (*dashed line*). **Top** the whole picture up to $n = 50$. **Bottom** a zoom of the left-bottom corner.

$$\left\langle k^2 \right\rangle^{\text{quasi}} = \frac{K_2^{\text{quasi}}}{n} = n - 3 + \frac{6}{n}. \tag{11}$$

Eq. 4 and Eq. 11 allow one to infer that

$$E_0[C] = \frac{n}{3} - 1 \tag{12}$$

for a quasi-star tree. Fig. 3 shows the linear growth of the expected number of crossings as a function of the number of vertices in quasi-star trees. Interestingly, if a tree has a value of $\left\langle k^2 \right\rangle$ that exceeds $\left\langle k^2 \right\rangle^{\text{quasi}}$, it has to be a star tree (see Appendix). For this reason, Fig. 3 suggests that star trees are the only option to obtain a small constant number of crossings. A detailed mathematical argument will be presented next.

If it is required that the expected number of crossings does not exceed a, i.e. $E_0[C] \leq a$, Eq. 4 gives

$$\left\langle k^2 \right\rangle \geq n - 1 - \frac{6a}{n}. \tag{13}$$

Notice that the preceding result has been derived making no assumption about the tree topology. We aim to investigate when a $E_0[C] \leq a$ implies a star tree.

As a tree whose value of $\left\langle k^2 \right\rangle$ exceeds $\left\langle k^2 \right\rangle^{\text{quasi}}$ must be a star tree (Appendix), Eq. 13 indicates that if

$$n - 1 - \frac{6a}{n} > \left\langle k^2 \right\rangle^{\text{quasi}} \tag{14}$$

then $E_0[C] \leq a$ requires a star tree. Applying the definition of $\left\langle k^2 \right\rangle^{\text{quasi}}$ in Eq. 11 to Eq. 14, we obtain that a star tree is needed to expect at most a crossings if

$$n > 3a + 3. \tag{15}$$

Thus, Eq. 15 implies that a hub tree is needed to expect at most one crossings by chance ($a = 1$) for $n > 6$ (this can be checked with the help of Fig. 3). In order to have at most one crossing by chance, the structural diversity must be minimum because star trees are the only possible labeled trees. To understand the heavy constraints imposed by $a = 1$ on the possible trees, consider $t(n)$, the number of unlabeled trees of n that can be formed (Table 1). When $n = 4$, the only trees than can be formed are a star tree and a linear tree, which gives $t(4) = 2$. In contrast, the star tree is only one out of 19320 possible unlabeled trees when $n = 16$. The decrease in diversity is more radical as n increases (Table 1). The choice of $a = 2$ does not change the scenario so much: Eq. 15 predicts that sentences of length $n > 9$ should have a star tree structure if no more than two crossings are to be expected. Real syntactic dependency trees from sufficiently longer sentences are far from star trees, e.g., Fig. 1 [32, 22, 21].

Table 1 $t(n)$, the number of unlabeled trees of n vertices [44].

n	$t(n)$
1	1
2	1
3	1
4	2
5	3
6	6
7	11
8	23
9	47
10	106
11	235
12	551
13	1301
14	3159
15	7741
16	19320
17	48629
18	123867
19	317955

4 Alternative Hypotheses

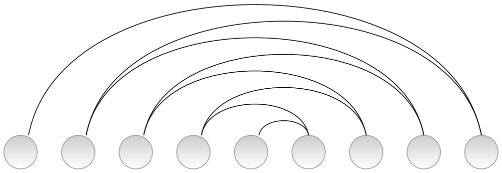
It has been shown that the low frequency of crossings is unexpected by chance in random linear arrangements of real syntactic dependency trees. As scientists, the next step is exploring the implications of this test and evaluating alternative hypotheses. Vertex degrees ($\langle k^2 \rangle$), which are an aspect of sentence structure, have been discarded as the only origin for the low frequency of crossings. This is relevant for some views where competence or grammar concern the structure of a sentence [3, 4, 6]. Discussing what competence or grammar is or should be is beyond the scope of this article but it is worth examining common reactions of language researchers when encountering a pattern:

- For statistical patterns such as Zipf's law for word frequencies and Menzerath's law, it was concluded that the patterns are inevitable [2, 45] (see [28, 46] for a review of the weaknesses of such conclusions).
- Concerning syntactic regularities in general, a naive but widely adopted approach is blaming (universal) grammar, the faculty of language or similar concepts [6, 15]. The fact that one is unable to explain a certain phenomenon through usage is considered as a justification for grammar (e.g., [6]). However, a rigorous justification requires a proof of the impossibility of usage to account for the phenomenon. To our knowledge, that proof is never provided.
- In the context of dependency grammar, crossings dependencies have been banned [31] or it has been argued that most phrases cannot have crossings or that crossings turn sentences ungrammatical [16, pp. 130]. It is worrying that the statement

is not the conclusion of a proof of the impossibility of a functional explanation. Furthermore, the argument of ”ungrammaticality” is circular and sweeps processing difficulties under the carpet.

In the traditional view of grammar or the faculty of language, the limited resources of the human brain are secondary or anecdotal [15, 16]. Recurring to grammar or a language faculty implies more assumptions, e.g. grammar would be the only reason why dependencies do not cross so often, and an explanation about the origins of the property would be left open (the explanation would be potentially incomplete). The property might have originated in grammar as a kind of inevitable logical or mathematical property, or might be supported by genetic information of our species, or it might also have been transferred to grammar (culturally or genetically) and so on. Thus, a grammar that is responsible for non-crossing dependencies would not be truly minimalistic (parsimonious) at all if the phenomenon could be explained by a universal principle of economy (universal in the sense of concerning the human brain not necessarily exclusively). This is likely the case of current approaches to dependency grammar at least (e.g. [16, 32, 31].

Fig. 4 A linear arrangements of the vertices of a linear tree that maximizes D (the sum of dependency lengths) when edge crossings are not allowed.



4.1 A Principle of Minimization of Dependency Crossings

A tempting hypothesis is a principle of minimization of dependency crossings (e.g., [47]) which can be seen as a quantitative implementation of the ban of crossings [31, 16]. This minimization can be understood as a purely formal principle (a principle of grammar detached from performance constraints but then problematic for the reasons explained above) or a principle related to performance. A principle of the minimization of crossings (or similar ones) is potentially problematic for at least three reasons:

- It is an inductive solution that may overfit the data.
- It is naive and superficial because it does not distinguish the consequences (e.g., uncrossing dependencies) from the causes. A deep theory of language requires distinguishing underlying principles from products: the principle of compression [24] from the law of abbreviation (a product), the principle of mutual information maximization from vocabulary learning biases [48] (another product), and so on.

- If patterns are directly translated into principles, the risk is that of constructing a fat theory of language when merging the tentative theories from every domain. When integrating the principle of minimization of crossings with a general theory of syntax, one may get two principles: a principle of minimization of crossings and a principle of dependency length minimization. In contrast, a theory where the minimization of crossings is seen as a side-effect of a principle of dependency length minimization [49, 47, 50, 39] might solve the problem in one shot through a single principle of dependency length minimization. However, it has cleverly been argued that a principle of minimization of crossings might imply a principle of dependency length minimization [47] and thus a principle of minimization of crossings might not imply any redundancy.

Thus, it is important to review the hypothesis of the minimization of dependency length and the logical and statistical relationship with the minimization of C.

4.2 A Principle of Minimization of Dependency Lengths

The length of a dependency is usually defined as the absolute difference between the positions involved (the 1st word of the sentence has position 1, the 2nd has position 2 and so on). In Fig. 1, the length of the dependency between ”John” and ”saw” is $2-1=1$ and the length of the dependency between ”saw” and ”dog” is $4-2=2$. In this definition, the units of length are word tokens (it might be more precise if defined in phonemes, for instance). If d_i is the length of the i-th dependency of a tree (there are $n-1$ dependencies in a tree) and $g(d)$ is the cost of a dependency of length d, the total cost of a linguistic sequence from the perspective of dependency length is the total sum of dependency costs [51, 52], which can defined as

$$D = \sum_{i=1}^{n} g(d_i), \tag{16}$$

where $g(d)$ is assumed to be a strictly monotonically increasing function of d [52]. The mean cost of a tree structure is defined as $\langle d \rangle = D/(n-1)$. If g is the identity function ($g(d) = d$) then D is the sum of dependency lengths (and $\langle d \rangle$ is the mean dependency length). It has been hypothesized that D or equivalently $\langle d \rangle$ is minimized by sentences (see [52] for a review). The hypothesis does not imply that the actual value of D has to be the minimum in absolute terms. Hereafter we assume that $g(d)$ is the identity function ($g(d) = d$).

The minimum D that can be obtained without altering the structure of the tree is the solution of the minimum linear arrangement problem in computer science [53]. Another baseline is provided by the expected value of D in a random arrangement of the vertices, which is [25]

$$E_0[D] = (n-1)(n+1)/3. \tag{17}$$

Statistical analyzes of D in real syntactic dependency trees have revealed that D is systematically below chance (below $E_0[D]$) for sufficiently long sentences but above the value of a minimum linear arrangement on average [25, 54].

D is one example of a metric or score to evaluate the efficiency of a sentence from a certain dimension (see [55, 56] for similar metrics on syntactic structures). Stating clearly the metric that is being optimized is a requirement for a rigorous claim about efficiency of language. For instance, consider the sentence on top of Fig. 5 and the version below that arises from the right-extraposition of the clause ”who I knew”. Notice that the dependency tree is the same in both cases (only word order varies). It has been argued that theories of processing based on the distance between dependents *”predict that an extraposed relative clause would be more difficult to process than an in situ, adjacent relative clause”* [57]. However, that does not grant one to conclude that the sentence on top of Fig. 5 should be easier to process than the sentence below from that perspective: one has $D = 3 \times 1 + 2 + 4 + 6 = 15$ for the sentence without the extraposition and $D = 3 \times 1 + 2 \times 2 + 3 = 10$ for the one with right-extraposition suggesting that the easier sentence is precisely the sentence with right-extraposition. The prediction about the cost of extraposition in [57] is an incomplete argument. The ultimate conclusion about the cost of extraposition requires considering all the dependency lengths, i.e. a true efficiency score. A score of sentence locality is needed to not rule out prematurely accounts of the processing difficulty of non-projective orderings that are based purely on *”dependency locality in terms of linear positioning”* [57]. The issue is tricky even for studies where a quantitative metric of dependency length such as D is employed: it is important to not mix values of the metric coming from sentences of different length to draw solid conclusions about a corpus or a language [54]. The need of strengthening quantitative standards [58] and also the need of appropriate controls [59, 26, 27] in linguistic research are challenges that require the serious commitment of each of us.

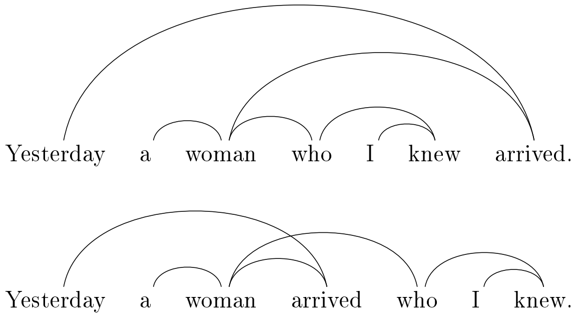

Yesterday a woman who I knew arrived.
Yesterday a woman arrived who I knew.

Fig. 5 **Top** an English sentence with a relative clause (”who I knew”). **Bottom** the same sentence with a right extraposition of the relative clause. Adapted from [57].

4.3 The Relationship Between Minimization of Crossings and Minimization of Dependency Lengths

Let us examine the logical relationship between the two principles above from the perspective of their global minima. On the one hand, the minimum value of C is 0 [39] and the minimum value of D is obtained by solving the minimum linear arrangement problem [60] or a generalization [52], which yields $D_{\min}$. At constant n and $\left\langle k^2 \right\rangle$, there are two facts:

- $C = 0$ does not imply $D = D_{\min}$ in general. This can be shown by means of two extreme configurations, a star tree, which maximizes $\left\langle k^2 \right\rangle$ and a linear tree, which minimizes $\left\langle k^2 \right\rangle$ [38]:
  - A star tree implies $C = 0$ [39]. In that tree, $D = D_{\min}$ holds only when the hub is placed at the center [52]. If the hub is placed at one of the extremes of the sequence, D is maximized for that tree [52]. Those results still hold when $g(d)$ is not the identity function but a strictly monotonically increasing function of d [52]. Furthermore, the placement of the hub in one extreme implies the maximum D that a non-crossing tree (not necessarily a star) can achieve, which is $D = n(n-1)/2$ [39].
  - A linear tree can be arranged linearly with $C = 0$ and $D = D_{\min} = n - 1$ (as in Fig. 2, which has no crossings and coincides with the smallest D than an unrestricted tree can achieve (as $d_i \geq 1$ and a tree has $n-1$ edges). In contrast, a linear arrangement of the kind of Fig. 4 has $C = 0$ but yields $D = n(n-1)/2$, i.e. the maximum value of D that a non-crossing tree can achieve [39].
- $D = D_{\min}$ does not imply $C = 0$ in general. It has been shown that a linear arrangement of vertices with crossings can achieve a smaller value of D than that of a minimum linear arrangement that minimizes D when no crossings are allowed [61] (Fig. 6).

Thus, there is not a clear relationship between the minima of D and C when one abstracts from the structural properties of real syntactic dependency trees. The impact of the real properties of dependency structures for the arguments should be investigated in the future.

As for the statistical relationship between C and D, statistical analyzes support the hypothesis of a positive correlation between both at least in the domain between $n-1$, the minimum value of D, and $D = E_0[D]$ [49, 38]. For instance, crossings practically disappear if the vertices of a random tree are ordered to minimize D. The relationship between C and D in random permutations of vertices of the dependency trees is illustrated in Fig. 7: C tends to increase as D increases from $D = D_{\min}$ onwards. Results obtained with similar metrics [47, 62] are consistent with such a correlation. For instance, a measure of dependency length reduces in random trees when crossings are disallowed [62].

This opens the problem of causality, namely if the minimization of D may cause a minimization of C, or a minimization of C may cause a minimization of D, or both

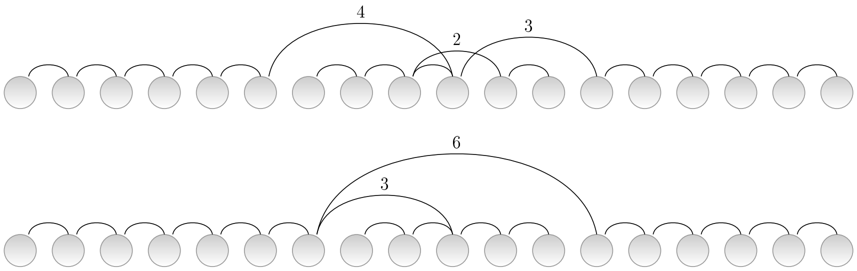


Fig. 6 Minimum linear arrangements of the same tree (only the length of edges that are longer than unity is indicated). **Top** a minimum linear arrangement of a tree. The total sum of dependency lengths is $D = 4+2+3+14 = 23$. **Bottom** a minimum linear arrangement of the same tree when crossings are disallowed. The total sum of dependency lengths is $D = 6+3+15 = 24$. Adapted from [61].

principles cannot be disentangled or simply both principles are epiphenomena (correlation does not imply causality). Solving the problem of causality is beyond the scope of this article but we can however attempt to determine rationally which of the two forces, minimization of D or minimization of C, might be the primary force by means of qualitative version of information theoretic model selection [13]. The apparent tie between these two principles will be broken by the more limited explanatory power of the minimization of C. The point is focusing on the phenomena that a principle of minimization of C cannot illuminate. A challenge for that principle is the ordering of subject (S), object (O) and verb (V). The dependency structure of that triple is a star tree with the verb as the hub [52]. A tree of less than four vertices cannot have crossings [39]. Thus, $C = 0$ regardless of the ordering of the elements of the triple. Interestingly, the principle of minimization of C cannot explain why languages abandon SOV in favor of SVO [63]. In contrast, the attraction of the verb towards the center combined with the structure of the word order permutation space can explain it [52]. Another challenge for a principle of the minimization of C are the relative ordering of dependents of nominal heads in SVO orders, that have been argued to preclude regression to SOV [52]. To sum up, a single principle of minimization of C would compromise explanatory power and if its limitations were complemented with an additional principle of dependency length minimization then parsimony would be compromised.

The reminder of the article is aimed at investigating a more parsimonious explanation for the ubiquity of non-crossing dependencies based on the minimization of D as the primary force [49, 47, 50, 39]. The minimization of D would be a consequence of the minimization of cognitive effort: longer dependencies are cognitively more expensive [47, 55, 64, 65]. We will investigate two null hypotheses that allow one to predict the number of crossings as function of dependency lengths, which are determined by cognitive pressures.

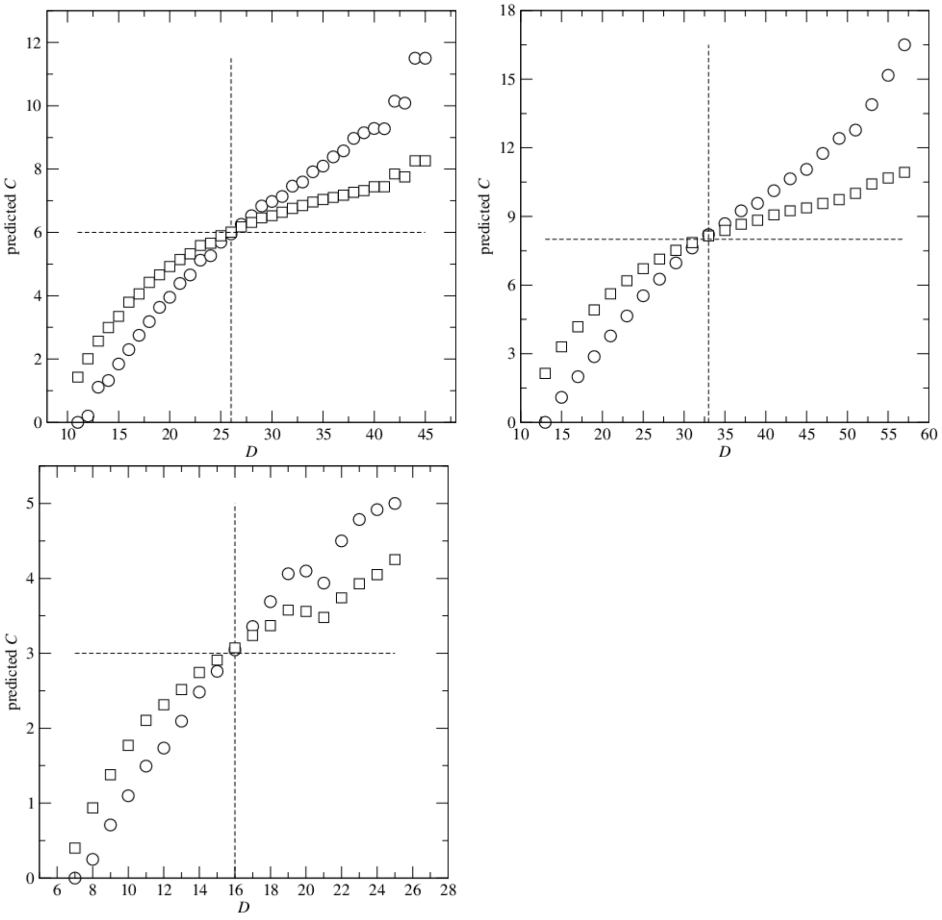


Fig. 7 Predictions about number of dependency crossings (C) as a function of the sum of dependency lengths (D) for dependency trees of real sentences. $E[C|D]$, the average C in all the possible permutations with a given value of D (*circles*), is compared against the average $E_1[C]$ in those permutations (*squares*). $E_1[C]$ is a prediction about C based on information on the distance of just one of the vertices potentially involved in a crossing. The *vertical dashed line* indicates the value of $E_0[D]$ and the *horizontal line* indicates the value of $E_0[C]$ (given a tree, those values are easy to compute with the help of Eqs. 17 and 4, respectively). The value of $E[C|D]$ and its prediction is only shown for values of D achieved by at least one permutation because certain values of D cannot be achieved by a given tree. For a given tree, $D_{\min}$ and $D_{\max}$ are, respectively, the minimum and the maximum value of D that can be reached ($D_{\min} \leq D \leq D_{\max}$). **Top-Left** sentence on top Fig. 1 with $D_{\min} = 11$, $D_{\max} = 45$, $E_0[D] = 26$ and $E_0[C] = 6$. **Top-Right** sentence at the bottom of Fig. 1 with $D_{\min} = 13$, $D_{\max} = 57$, $E_0[D] = 33$ and $E_0[C] = 8$. Even values of D are not found. **Bottom-Left** results for the two sentences in Fig. 5 with $D_{\min} = 7$, $D_{\max} = 25$, $E_0[D] = 16$ and $E_0[C] = 3$. Notice that the results are valid for the couple of sentences in Fig. 5 because they have the same structure in a different order. For this reason it is not surprising that $D_{\min}$, $D_{\max}$, $E_0[D]$, $E_0[C]$ and $E[C|D]$ coincide. Interestingly, $E_1[C]$ turns out to be the same for both, too.

5 A Stronger Null Hypothesis

Here we consider a predictor for the number of crossings when some information about the length of the arcs is known. The predictor guesses that number by considering, for every pair of edges that may potentially cross (pairs of edges sharing vertices cannot cross), the probability that they cross knowing the length of one of the edges and assuming that the vertices of the other edge have been arranged linearly at random. The null hypothesis in Sect. 3 predicts the number of crossings in the same fashion but replacing that probability by the probability that two edges cross when both are arranged linearly at random (not arc length is given).

The null hypothesis in Sect. 3 and the null hypothesis that will be explored in the current section, are reminiscent of two null hypotheses that are used in networks theory: random binomial graphs and random graphs with an arbitrary degree sequence or degree distribution [66, 67, 68] (in our case, information about dependency length plays an equivalent role to vertex degree in those models).

5.1 The Probability that Two Edges Cross

Let $\pi(v)$ be the position of the vertex v in the linear arrangement ($\pi(v) = 1$ if v is the 1st vertex of the sequence, $\pi(v) = 2$ it is the 2nd vertex and so on). $u \sim v$ is used to indicate an edge between vertices u and v, namely that u and v are connected. A vertex position q is covered by the edge $u \sim v$ if and only if $min(\pi(u), \pi(v)) < q < max(\pi(u), \pi(v))$. A position q is external to the edge $u \sim v$ if and only if $q < min(\pi(u), \pi(v))$ or $q > max(\pi(u), \pi(v))$. $s \sim t$ crosses $u \sim v$ if and only if

- either $\pi(s)$ is covered by $u \sim v$ and $\pi(t)$ is external to $u \sim v$
- or $\pi(t)$ is covered by $u \sim v$ and $\pi(s)$ is external to $u \sim v$.

Notice that edges that share vertices cannot cross.

Let us consider first that no information is known about the length of two arcs. The probability that two edges, $s \sim t$ and $u \sim v$, cross when arranged linearly at random is $p(\text{cross}) = 1/3$ if the edges do not share any vertex and $p(\text{cross}) = 0$ otherwise [38]. We will investigate $p(\text{cross}|d)$, the probability that two edges, $s \sim t$ and $u \sim v$, cross when arranged linearly at random knowing that (a) one of the edges has length d, e.g., $|\pi(u) - \pi(v)| = d$ and (b) the edges do not share any vertex.

If $s \sim t$ and $u \sim v$ share a vertex, then $p(\text{cross}|d) = 0$. If not,

Proposition 2.

$$\begin{aligned} p(cross|d) &= 2\frac{(d-1)(n-d-1)}{(n-2)(n-3)} \\ &= \frac{2(-d^2+nd-n+1)}{(n-2)(n-3)}. \end{aligned} \tag{18}$$

Proof. To see this notice that $d-1$ is the number of vertex positions covered by the edge of length d and $n-d-1$ is the number of vertices that are external to that edge. Once the edge of length d has been arranged linearly, there are

$$\binom{n-2}{2} = \frac{(n-2)(n-3)}{2} \tag{19}$$

possible placements for the two vertices of the other edge of which only $(d-1)(n-d-1)$ involve a position covered by the edge of length d and another one that is external to that edge. □

$p(\text{cross}|d)$ and $p(\text{cross}) = 1/3$ are related, i.e.

$$\sum_{d=1}^{n-1} p(\text{cross}|d)p(d) = \frac{1}{3}, \tag{20}$$

where

$$p(d) = \frac{2(n-d)}{n(n-1)} \tag{21}$$

is the probability that the linear arrangement of the two vertices of an edge yields a dependency of length d [25]. Eq. 20 is easy to prove applying the definition of conditional probability ($p(\text{cross}|d) = p(\text{cross}, d)/p(d)$), which gives

$$\sum_{d=1}^{n-1} p(\text{cross}|d)p(d) = \sum_{d=1}^{n-1} p(\text{cross}, d) = p(\text{cross}) = \frac{1}{3}. \tag{22}$$

When n takes the smallest value needed for the possibility of crossings, i.e. $n=4$ [38], Eq. 18 yields $p(\text{cross}|1) = p(\text{cross}|3) = 0$ and $p(\text{cross}|2) = 1$. It is easy to show that

- $p(\text{cross}|d)$ is symmetric, i.e. $p(\text{cross}|d) = p(\text{cross}|n-d)$,
- $p(\text{cross}|d)$ has two minima ($p(\text{cross}|d) = 0$), at $d=1$ and $d=n-1$.
- $$p(\text{cross}|d) \leq p_{\max}(\text{cross}|d), \tag{23}$$

  where

  $$p_{\max}(\text{cross}|d) = \frac{\frac{n^2}{2} - 2(n-1)}{(n-2)(n-3)}. \tag{24}$$

  To see this notice that $p(\text{cross}|d)$ is a function (Eq. 18) that has a maximum at $d = d^* = n/2$. Applying $d = d^*$, Eq. 18 gives Eq. 24. As $p(\text{cross}|d)$ is not defined for non-integer values of d, equality in Eq. 23 needs that d^* is integer, namely that n is even. In the limit of large n, one has that $p_{\max}(\text{cross}|d) = 1/2$.
- Accordingly, $p(\text{cross}|d)$ has either a maximum at $d = n/2$ if n is even or two maxima, at $d = \lfloor n/2 \rfloor$ and $d = \lceil n/2 \rceil$ when n is even because d is a natural number.

5.2 The Expected Number of Edge Crossings

Imagine that the structure of the tree is defined by an adjacency matrix $A = \{a_{uv}\}$ such that $a_{uv} = 1$ if the vertices u and v are linked and $a_{uv} = 0$ otherwise. Let C be the number of edge crossings and $C(u,v)$ be the number of crossings where the edge formed by u and v is involved ($C(u,v) = 0$ if u and v are unlinked), i.e.

$$C = \frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}C(u,v) \tag{25}$$

and

$$C(u,v) = \frac{1}{2}\sum_{s=1,s\neq u,v}^{n}\;\sum_{t=1,t\neq u,v}^{n} a_{st}C(u,v;s,t), \tag{26}$$

where $C(u,v;s,t)$ indicates if u,v and s,t define a couple of edges that cross ($C(u,v;s,t) = 1$ if they cross, $C(u,v;s,t) = 0$ otherwise). Thus, the expectation of C is

$$E[C] = \frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}E[C(u,v)]. \tag{27}$$

In turn, the expectation of $C(u,v)$ is

$$E[C(u,v)] = \frac{1}{2}\sum_{s=1,s\neq u,v}^{n}\;\sum_{t=1,t\neq u,v}^{n} a_{st}E[C(u,v;s,t)]. \tag{28}$$

As $C(u,v;s,t)$ is an indicator variable,

$$E[C(u,v;s,t)] = p(s \sim t \ \&\ u \sim v \text{ cross}), \tag{29}$$

namely $E[C(u,v;s,t)]$ is the probability that the edges $s \sim t$ and $u \sim v$ cross knowing that they do not share any vertex.

The number of edges that can cross the edge $u \sim v$ is $n - k_u - k_v$ (edges that share a vertex with $u \sim v$ cannot cross), which gives [38]

$$E[C(u,v)] = (n - k_u - k_v)p(\text{cross}|u \sim v) \tag{30}$$

assuming that the probability that the edge $u \sim v$ crosses another edge depends at most on $u \sim v$. Applying the last result to Eq. 27, we obtain a general predictor of the number of crossings under a null of hypothesis x, i.e.

$$E_x[C] = \frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}(n - k_u - k_v)p_x(\text{cross}|u \sim v), \tag{31}$$

where x indicates if the identity of one of the edges potentially involved a crossing, i.e. $u \sim v$ is actually known ($x = 1$ if it is known; $x = 0$ otherwise). Eq. 31 defines a family of predictors where x is the parameter. $x = 0$ corresponds to the null hypothesis in Sect. 3. To see it, notice that $x = 0$, i.e.

$$p_x(\text{cross}|u \sim v) = p(\text{cross}) = 1/3 \tag{32}$$

transforms Eq. 31 into Eq. 4 [38]. $E_x[C]$ can be seen as a simple approximation to the expected number of crossings in a linear random arrangement of vertices when all edge lengths are given. While $E_0[C]$ is a true expectation, $E_1[C]$ is not for not conditioning on the same lengths for every pair of edges that may cross.

A potentially more accurate prediction of C with regard to Eq. 4 is obtained when $x = 1$. For simplicity, let us reduce the knowledge of an edge to the knowledge of its length. Let us define $d_{uv} = |\pi(u) - \pi(v)|$ as the distance between the vertices u and v. If $x = 1$, the substitution of $p_x(\text{cross}|u \sim v)$ by $p(\text{cross}|d_{uv})$ in Eq. 31 yields $E_1[C]$, a prediction of C where, for ever pair of edges that many potentially cross, the length of one edge is given and a random placement is assumed for the other edge, i.e.

$$E_1[C] = \frac{1}{4}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}(n - k_u - k_v)p(\text{cross}|d_{uv}). \tag{33}$$

Fig. 7 shows that $E_1[C]$ is positively correlated with D on permutations of the vertices of the trees of a real sentences. Interestingly, the values of $E_1[C]$ overestimate the average C when $D < E_0[D]$ and underestimate it when $D > E_0[D]$.

Variations in dependency lengths can alter $E_1[C]$. A drop in $p(\text{cross}|d_{uv})$ leads to a drop in $E_1[C]$. It has been shown above that $p(\text{cross}|d_{uv})$ is minimized by $d = 1$ and $d = n-1$ and maximized by $d_{uv} \approx n/2$. When $d_{uv} < n/2$, a decrease in d_{uv} decreases $p(\text{cross}|d_{uv})$. In contrast, a decrease in d_{uv} when $d_{uv} > n/2$, increases $p(\text{cross}|d_{uv})$. However, the shortening of edges is more likely to decrease $p(\text{cross}|d)$ because

- Tentatively, the linear arrangement of a tree can only have $n-d$ edges of length d [25].
- Under the null hypothesis that edges are arranged linearly at random, short edges are more likely than long edges. In that case, $p(d)$, the probability that an edge has length d satisfies $p(d) \propto n-d$ (Eq. 21).
- The potential ordering of a sentence is unlikely to have many dependencies of length greater than about $n/2$ because the cost of a dependency is positively correlated with its length [55, 69].
- From an evolutionary perspective, initial states are unlikely to involve many long dependencies [51].

Thus, it is unlikely that the shortening of edges increases $E_1[C]$. Instead, this is likely to decrease $E_1[C]$. Fig. 7 shows that

- the average $E_1[C]$ is tends to increase as D increases
- the mean C is bounded above by $E_1[C]$ (in the domain $D \leq E_0[D]$)

in concrete sentences.

Applying the definition of $p(\text{cross}|d)$, given in Eq. 18, to Eq. 33, yields

$$E_1[C] = \frac{B_2 - B_1}{(n-2)(n-3)}, \tag{34}$$

where

$$B_2 = \frac{1}{2}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}(n-k_u-k_v)(n-d_{uv})d_{uv} \tag{35}$$

and

$$\begin{aligned}
B_1 &= \frac{n-1}{2}\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv}(n-k_u-k_v) \\
&= \frac{n-1}{2}\left(n\sum_{u=1}^{n}\sum_{v=1}^{n} a_{uv} - 2\sum_{u=1}^{n} k_{uv}\sum_{v=1}^{n} a_{uv}\right) \\
&= \frac{n-1}{2}\left(n\sum_{u=1}^{n} k_u - 2\sum_{u=1}^{n} k_u^2\right) \\
&= \frac{n-1}{2}\left(2n(n-1) - 2n\left\langle k^2\right\rangle\right) && \text{applying Eqs. 2 and 5} \\
&= n(n-1)\left(n-1-\left\langle k^2\right\rangle\right). && (36)
\end{aligned}$$

On the one hand, notice that $B_1 \geq 0$ because $n \geq 1$ and $\left\langle k^2\right\rangle \leq n-1$ [39]. On the other hand, notice that $B_2 \geq 0$ because

- The fact that $C(u,v) < 0$ is impossible by definition and also the fact that $C(u,v) \leq n-k_u-k_v$ [39] yields $n-k_u-k_v \geq 0$.
- $d_{uv} \leq n-1$ by definition [39].

Thus, $E_1[C]$ is proportional to the difference between two terms: a term B_2 that depends on dependency lengths and a term B_1 that depends exclusively on vertex degrees and sentence length (in words). Interestingly, Eq. 4 allows one to see $E_1[C]$ as a function of $E_0[C]$ since $B_1 = 6(n-1)E_0[C]$.

6 Another Stronger Null Hypothesis

Another prediction from a stronger null hypothesis is $E[C|D]$, the expected value of C when one of the orderings (permutations) preserving the actual value of D is chosen uniformly at random. Fig. 7 shows that knowing the exact value of D, a positive correlation between $E[C|D]$ and D follows for a concrete sentence (this is what circles are showing) and, interestingly, $E[C|D]$ predicts a smaller number of crossings than the average $E_1[C]$ as a function of D.

7 Predictions, Testing and Selection

Table 2 compares $E_0[C]$, $E_1[C]$ and $E[C|D]$ for the example sentences that have appeared so far. Table 2 shows that, according to the normalized error,

- $E_0[C]$ makes the worst predictions.
- The predictions of $E[C|D]$ are the best except in one sentence where $E_1[C]$ wins.

The take home message here is that it is possible to make quantitative predictions about the number of crossings, an aspect that theoretical approaches to the problem of crossing dependencies have missed [50, 70, 57]. This quantitative requirement should not be regarded as a mathematical *divertimento*: science is about predictions [71].

Table 2 The predicted crossings by each of the three null hypotheses, i.e. (1) random linear arrangement of vertices, (2) random linear arrangement with knowledge of the length of one of the edges that can potentially cross, and (3) random linear arrangement of vertices at constant sum of dependency lengths, for the example sentences employed in this article. Results of p-value testing for the 3rd null hypothesis are also included. n is the number of vertices of the tree, $\langle k^2 \rangle$ is the degree second moment about zero, D is the sum of dependency lengths, C is the actual number of crossings, $C_{\max}$ is the potential number of crossings, i.e. $C_{\max} = n\left(n-1-\langle k^2 \rangle\right)/2$ [39]. $E_0[C]$, $E_1[C]$ and $E[C|D]$ are the predicted number of crossings according to the 1st, 2nd and 3rd hypotheses, respectively. $\varepsilon_0[C]$, $\varepsilon_1[C]$ and $\varepsilon[C|D]$ are the normalized error of the 1st, 2nd and 3rd hypotheses, respectively, i.e. $\varepsilon_x[C] = |C - E_x[C]|/C_{\max}$ and $\varepsilon[C|D] = |C - E[C|D]|/C_{\max}$. Left and right p-values are provided for two statistics, C and $|C - E[C|D]|$ under the 3rd null hypothesis. R is the number of permutations of the vertex sequence where D coincides with the original value.

	Fig. 1, top	Fig. 1, bottom	Fig. 5, top	Fig. 5, bottom
n	9	10	7	7
$\langle k^2 \rangle$	4	4.2	3.4	3.4
D	13	17	15	10
C	0	1	0	1
$C_{\max}$	18	24	9	9
$E_0[C]$	6	8	3	3
$\varepsilon_0[C]$	0.33	0.29	0.33	0.22
$E_1[C]$	2.4	4.2	1.2	2.2
$\varepsilon_1[C]$	0.13	0.14	0.13	0.13
$E[C|D]$	1.1	2	2.8	1.1
$\varepsilon[C|D]$	0.062	0.041	0.31	0.011
Left p-value of C	0.28	0.37	0.058	0.69
Right p-value of C	1	0.88	1	0.75
Left p-value of $|C - E[C|D]|$	0.94	0.54	0.97	0.43
Right p-value of $|C - E[C|D]|$	0.33	0.71	0.088	1
R	288	6664	548	102

Our exploration of some sentences suggest that $E[C|D]$ makes better predictions in general. Notwithstanding we cannot rush to claim that $E[C|D]$ is the best model among those that we have considered only for that reason. According to standard model selection, the best model is one with the best compromise between the quality of its fit (the error of its predictions) and parsimony (the number of parameters) [13]. Applying a rigorous framework for model selection is beyond the scope of this article but we can examine the parsimony of each model qualitatively to shed light on the best model. Interestingly, the three predictors vary concerning the amount of information that suffices to make a prediction. A comparison of the minimal information that each needs to make a prediction would be a better approach but that is beyond the scope of the current article. The value of $E_0[C]$ can be computed knowing only n and $\langle k^2 \rangle$ (Eq. 4). The value of $E_1[C]$ can be computed knowing only n, the length of every edge and the degrees of the nodes forming every edge (Eqs. 18 and 33). The calculation of $E[C|D]$ is less demanding than that of $E_1[C]$ concerning edge lengths. i.e. the total sum of the their lengths suffices (the length of individual edges is irrelevant), but still employs some information about edges, e.g., the nodes forming every edge (Sect. 6). Our preliminary results (Table 2) and our analysis of parsimony from the perspective of sufficient information suggests that $E[C|D]$ has a better trade-off between the quality of its predictions and parsimony with respect to $E_1[C]$. Interestingly, none of the models has free parameters if the ordering of the vertices and the syntactic dependency structure of the sentence is known as we have assumed so far.

It is possible to perform traditional p-value testing of our models. For simplicity we focus on the null hypothesis that is defined in Sect. 6, i.e. permutations of vertices where the sum of edge lengths coincides with the true value (Table 2 shows the number of permutations of this kind for each sentence). We consider two statistics for test test. First, C, the actual number of crossings of a dependency tree. Then one can define the left p-value as the proportion of those permutations with a number of crossings at most as large as the true value. Similarly, one can define the right p-value as the proportion of those permutations with a number of crossings at least as large as the true value. One has to choose a significance level α. The significance level must be such that there can be p-values bounded above by α a priori. Otherwise, one is condemned to make type II errors. The smallest possible p-value is $1/R$, where R is the number of permutations yielding the original value of D (there is at least one permutation giving the same value of the statistic, i.e. the one that coincides with the original linear arrangement). Thus, α must exceed $1/R_{\min}$, being $R_{\min}$ the smallest value of R in Table 2. $R_{\min} = 102$ yields $\alpha \geq 1/102 \approx 0.01$. Thus we can safely choose a significance level of $\alpha = 0.05$. Table 2 shows that the left and right p-values are always below α, suggesting that the real numbers of crossings are compatible with those of this null hypothesis. However, notice that the p-value is borderline for one sentence (Fig. 5, top). Second, we consider $|C - E[C|D]|$, the absolute value of the difference between the actual number of crossings and the expected value, as another statistic to perform p-value testing. Accordingly, we define left and right p-values for the latter statistic following the same rationale used for the p-values of C. Table 2 shows $|C - E[C|D]|$ is neither significantly low nor significantly large, thus providing support for the hypothesis that the number of crossings is determined by global constraints on dependency lengths.

Our p-value tests should be seen as preliminary statistical attempts. Future research should involve more languages and large dependency treebanks. Moreover, information theoretic models selection offers a much powerful approach over traditional p-value testing [13] and should be explored in the future.

8 Discussion

In this article, the possibility that the low number of crossings in dependency trees is a mere consequence of chance has been considered. As the expected number of crossings (when edge lengths are not given) decreases as the degree 2nd moment increases, a high hubiness could lead to a small number of crossings by chance, when dependency lengths are unconstrained. However, it has been shown that the hubiness required to have a small number of crossings in that circumstance would imply star trees, which are problematic for at least two reasons: real syntactic dependency trees of a sufficient length are not star trees and the diversity of possible syntactic dependency structures would be seriously compromised (Sect. 3). One cannot exclude that hubiness has some role in decreasing the potential number of crossings in sentences [39, 62] but it cannot be the only reason. ”Grammar” has been examined as an explanation for the rather low frequency of crossings and the more parsimonious hypothesis of the minimization of syntactic dependency lengths [49, 47, 50, 39] has been revisited (Sect. 4). Stronger null hypotheses involving partial information about dependency lengths suggest that the shortening of the dependencies is likely to imply a reduction of crossings (recall empirical evidence in Fig. 7 and general mathematical arguments based on one of those stronger null hypotheses in Sect. 5.2). Moreover, it has been shown that a null hypothesis incorporating global information about dependency lengths, the sum of dependency lengths (Sect. 6) allows one to make specially accurate predictions about the actual number of crossings. The error of those predictions is neither surprisingly low nor high (Sect. 7). Our findings provide support for the hypothesis that uncrossing dependencies could be a side-effect of dependency length minimization, a principle that derives from the limited computational resources of the human brain [47, 55, 64, 65]. A universal grammar, a faculty of language or a competence-plus limiting the number of crossings might not be necessary. Just a version of Zipf’s least effort might suffice [72].

Upon a superficial analysis of facts, it is tempting to conclude that crossings cause processing difficulties and thus should be reduced. That follows easily from the correlation between crossings and dependency lengths that has been found in real and artificial sentence structures (e.g., [38, 62]). However, the cognitive cost of crossings does not need to be a direct consequence of the crossing [47] but a side effect of the longer dependencies that crossings are likely to involve [49, 50]. It has been argued that a single principle of minimization of dependency crossings would compromise seriously parsimony and explanatory power (Sect. 4).

Although Fig. 7 shows that solving the minimum linear arrangement problem yields zero crossings for concrete sentences ($E[C|D] = 0$ and thus $C = 0$ for $D = D_{\text{min}}$), it is important to bear in mind that dependency length minimization cannot promise to reduce crossings to zero in all cases: the minimum linear arrangement of a tree can involve crossings (Fig. 6). Interestingly, this means that the presence of some crossings in a sentence does not contradict a priori pressure for dependency length minimization or pressure for efficiency. Recall also the examples of real English in Fig. 5, showing that an ordering without crossings can have a higher sum of dependency lengths than one with crossings. This is specially important for Dutch, where crossing structures abound and have been shown to be easier to process that parallel non-crossing structures in German [73]. We believe that our theoretical framework might help to illuminate experiments suggesting that orderings with crossings tax working memory less than orderings with nesting [70].

In this article, we have addressed the problem of non-crossing dependencies from a really theoretical perspective. The arguments are a priori valid for any language and any linguistic phenomenon. This is a totally different approach from the investigation of non-crossing dependencies in a given language with a specific phenomenon (e.g., extraposition of relative clauses in English as in [57], see also [73] for other languages). With such a narrow focus, the development of a general theory of word order is difficult. Generativists have been criticized for not having developed a general theory of language but a theory of English [74]. If one takes seriously recent concerns about the limits of building a theory from a sample of languages [75], it follows that hypotheses about non-crossing dependencies that abstract from linguistic details like ours (see also [70]) should receive more attention in the future.

Although we believe that non-crossing dependencies are the main reason why crossings dependencies do not occur very often in languages, we do not believe that cognitive pressure for dependency length minimization is the only factor involved in word order phenomena. The maximization of predictability or the structure of word order permutation space are crucial ingredients for a non-reductionist theory of word order [52, 51]. Word order is a multiconstraint engineering problem [51].

Tentatively, a deep theory of syntax does not imply grammar or a language faculty exclusively. A well-known example is the case of sentence acceptability that may derive in some cases from processing constraints [55, 8]. Our findings suggest that grammar may not be an autonomous entity but a series of phenomena emerging from physical or psychological constraints. Grammar might simply be an epiphenomenon [9]. This is a more parsimonious hypothesis than grammar as a conventionalization of processing constraints [8]. Grammar may require fewer parameters than commonly believed. This is consistent with the idea that it would be desirable that the quantitative constraints of competence-plus are replaced by a theory of processing complexity or that the content of the ”plus” derives from memory and integration costs [7]. Being the ”plus” part non-empty, the point is elucidating whether the ”competence” part is indeed empty or at least lighter than commonly believed.

A deep theory of language cannot be circumscribed to language: a deep physical theory for the fall of inanimate objects is also valid for animate objects (with certain refinements). A deep theory of syntactic structures and their linear arrangement does not need to be valid only for human language but also for other natural systems producing and processing sequences and operating under limited resources. For these reasons, our results should be extended to non-tree structures to investigate crossings in RNA structures [23].

Acknowledgements The deep and comprehensive review of Steffen Eger has been crucial for the preparation of the final version of this article. We owe him the current proof of Proposition 4. We are also grateful to D. Blasi, R. Čech, M. Christiansen, J. M. Fontana, S. Frank, E. Gibson, H.

Liu and G. Morrill for helpful comments or discussions. All remaining errors are entirely ours. This work was supported by the grant *Iniciació i reincorporació a la recerca* from the Universitat Politècnica de Catalunya and the grants BASMATI (TIN2011-27479-C04-03) and OpenMT-2 (TIN2009-14675-C03) from the Spanish Ministry of Science and Innovation.

Appendix

Tree Reduction

As any tree of at least two vertices has at least two leaves [33, p. 11], a tree of $n+1$ vertices (with $n > 1$) yields a reduced tree of n vertices by removing one of its leaves. Notice that such a reduction from a tree of $n+1$ to a tree of n nodes will never produce a disconnected graph.

Consider that a tree has a leaf that is attached to a vertex of degree k. Then, the sum of squared degrees of a tree of $n+1$ vertices, i.e. $K_2(n+1)$, can be expressed as a function of $K_2(n)$, the sum of squared degrees of a reduced tree of n vertices, i. e.

$$\begin{aligned} K_2(n+1) &= K_2(n)+k^2-(k-1)^2+1 \\ &= K_2(n)+2k \end{aligned} \tag{37}$$

with $n \geq 0$.

The Only Tree that Has Degree Second Moment Greater than that of a Quasi-star Tree is a Star Tree.

A quasi-star tree is a tree with one vertex of degree $n-2$, one vertex of degree 2 and the remainder of vertices of degree 1. As a tree must be connected, that tree needs $n > 2$. The sum of squared degrees of a quasi-star tree is

$$K_2^{\text{quasi}}(n) = (n-2)^2+4+n-2 = n^2-3n+6. \tag{38}$$

The sum of squared degrees of a star tree is [39]

$$K_2^{\text{star}}(n) = n(n-1). \tag{39}$$

$K_2^{\text{star}}(n)$ is an upper bound of $K_2^{\text{quasi}}(n)$, more precisely,

Proposition 3. *For all $n \geq 3$,*

$$K_2^{quasi}(n) \leq K_2^{star}(n) \tag{40}$$

with equality if and only if $n = 3$.

Proof. Applying the definitions in Eqs. 38 and 39 to Eq. 40 we obtain

$$n^2 - 3n + 6 \leq n(n-1), \tag{41}$$

that is $n \geq 3$. □

$K_2(n)$ is maximized by star trees [39]. Quasi-star trees yield the second largest possible value of $K_2(n)$, more precisely

Proposition 4. *For all $n \geq 3$, it holds that*

$$K_2(n) > K_2^{quasi}(n) \implies K_2(n) = K_2^{star}(n) \tag{42}$$

for any tree with n vertices.

Proof. Denote the antecedent of the implication in Eq. 42 by $L(n)$ and the consequent by $R(n)$. We show by induction that, for all $n \geq 3$, $L(n) \implies R(n)$.

For $n = 3$, the fact that the only possible tree is both a star and a quasi-star tree implies that $L(n)$ is false. Thus, Eq. 42 holds trivially.

Let $n > 3$. For the induction step, assume that $L(n) \implies R(n)$, and also assume that $L(n+1)$ holds. We must show that then also $R(n+1)$ holds.

Consider an arbitrary tree T_{n+1} with $n+1$ vertices and consider the tree T_n, on n vertices, with a leaf l removed from T_{n+1}.

If $L(n)$ holds, then, by the induction hypothesis, $R(n)$ holds, i.e., $K_2(n) = K_2^{\text{star}}(n)$. A tree with n vertices for which $R(n)$ holds must be a star tree [38]; thus, T_n is a star tree. Then, the leaf vertex l is

- either attached to the hub of the star tree, in which case the resulting tree T_{n+1} is also a star tree, so that $K_2(n+1) = K_2^{\text{star}}(n+1)$, i.e., $R(n+1)$, holds.
- or attached to a leaf of T_n, in which case T_{n+1} is a quasi-star tree, contradicting that $L(n+1)$ holds.

Conversely, if $L(n)$ does not hold, then $K_2(n) \leq K_2^{\text{quasi}}(n)$. Accordingly, a tree T_{n+1} with a leaf of degree k satisfies (for any k)

$$K_2(n+1) = K_2(n) + 2k \leq K_2^{\text{quasi}}(n) + 2k = n^2 - 3n + 6 + 2k, \tag{43}$$

being $K_2(n)$ the sum of squared degrees of the reduced tree. Now, we have assumed that $L(n+1)$ holds, i.e., that

$$K_2(n+1) > K_2^{\text{quasi}}(n+1) = n^2 - n + 4, \tag{44}$$

thanks to Eq. 38. Combining Eqs. 43 and 44 it is obtained

$$n^2 - n + 4 < K_2(n+1) \leq n^2 - 3n + 6 + 2k,$$

which implies

$$2n < 2 + 2k.$$

But this would require that $k > n-1$, that is, $k = n$, since the maximum degree of a vertex in a tree with $n+1$ vertices is n. In other words, T_{n+1} would be forced to have a vertex of degree $k = n$, whence T_{n+1} is a star tree, so that T_n also is a star tree (removing a leaf from a star tree yields a star tree). But, this would contradict that $L(n)$ does not hold since $K_2^{\text{star}}(n) > K_2^{\text{quasi}}(n)$ for $n > 3$ (Proposition 3). □

References

1. R. McDonald, F. Pereira, K. Ribarov, J. Hajič, in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Stroudsburg, PA, USA, 2005), HLT '05, pp. 523–530. DOI 10.3115/1220575.1220641
2. G.A. Miller, N. Chomsky, in *Handbook of Mathematical Psychology*, vol. 2, ed. by R.D. Luce, R. Bush, E. Galanter (Wiley, New York, 1963), pp. 419–491
3. N. Chomsky, *Aspects of the Theory of Syntax* (MIT Press, Cambridge, MA, 1965)
4. R. Jackendoff, *Foundations of Language* (Oxford University Press, Oxford, 1999)
5. F. Newmeyer, Journal of Linguistics **37**, 101 (2001). DOI 10.1017/S0022226701008593
6. F.J. Newmeyer, Language **79**, 682 (2003). DOI 10.1353/lan.2003.0260
7. J.R. Hurford, *Chapter 3. Syntax in the Light of Evolution* (Oxford University Press, Oxford, 2012), pp. 175–258
8. J.A. Hawkins, *Efficiency and Complexity in Grammars* (Oxford University Press, Oxford, 2004)
9. P.J. Hopper, in *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, ed. by M. Tomasello (Lawrence Erlbaum, Mahwah, NJ, 1998), pp. 155–175
10. R. Köhler, in *Quantitative Linguistik. Ein internationales Handbuch. Quantitative Linguistics: An International Handbook* (Walter de Gruyter, Berlin/New York, 2005), pp. 760–775
11. C.D. Manning, in *Probabilistic Linguistics*, ed. by R. Bod, J. Hay, S. Jannedy (MIT Press, Cambridge, MA, 2002), pp. 289–341
12. M.H. Christiansen, N. Chater, Cognitive Science **23**(2), 157 (1999). DOI 10.1207/s15516709cog2302_2
13. K.P. Burnham, D.R. Anderson, *Model Selection and Multimodel Inference. A Practical Information-theoretic Approach*, 2nd edn. (Springer, New York, 2002)
14. N. Chomsky, *The Minimalist Program* (MIT Press, 1995)
15. M.D. Hauser, N. Chomsky, W.T. Fitch, Science **298**, 1569 (2002). DOI 10.1126/science.298.5598.1569
16. R. Hudson, *Language Networks. The New Word Grammar* (Oxford University Press, Oxford, 2007)
17. S. Frank, R. Bod, M.H. Christiansen, Proceedings of the Royal Society B: Biological Sciences **279**, 45224531 (2012). DOI 10.1098/rspb.2012.1741
18. G.A. Miller, in *The Psycho-Biology of Language: an Introduction to Dynamic Psychology (by G. K. Zipf)* (MIT Press, Cambridge, MA, USA, 1968), pp. v–x
19. P. Niyogi, R.C. Berwick, A.I. Memo No. 1530 / C.B.C.L. Paper No. 118 (1995)
20. R. Suzuki, P.L. Tyack, J. Buck, Anim. Behav. **69**, 9 (2005). DOI 10.1016/j.anbehav.2004.08.004
21. Y. Lecerf, Rapport CETIS No. 4 pp. 1–24 (1960). Euratom
22. D. Hays, Language **40**, 511 (1964)
23. W.Y.C. Chen, H.S.W. Han, C.M. Reidys, Proceedings of the National Academy of Sciences **106**(52), 22061 (2009). DOI 10.1073/pnas.0907269106

24. R. Ferrer-i-Cancho, A. Hernández-Fernández, D. Lusseau, G. Agoramoorthy, M.J. Hsu, S. Semple, Cognitive Science **37**(8), 15651578 (2013). DOI 10.1111/cogs.12061
25. R. Ferrer-i-Cancho, Physical Review E **70**, 056135 (2004). DOI 10.1103/PhysRevE.70.056135
26. R. Ferrer-i-Cancho, F. Moscoso del Prado Martín, Journal of Statistical Mechanics p. L12002 (2011). DOI 10.1088/1742-5468/2013/07/L07001
27. F. Moscoso del Prado, in *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, ed. by M. Knauff, M. Pauen, N. Sebanz, I. Wachsmuth (Cognitive Science Society, Austin, TX, 2013), pp. 1032–1037
28. R. Ferrer-i-Cancho, B. Elvevåg, PLoS ONE **5**(4), e9411 (2009). DOI 10.1371/journal.pone.0009411
29. S.T. Piantadosi, H. Tily, E. Gibson, Proceedings of the National Academy of Sciences **108**(9), 3526 (2011)
30. J. Uriagereka, *Rhyme and Reason. An Introduction to Minimalist Syntax* (The MIT Press, Cambridge, Massachusetts, 1998)
31. R. Hudson, *Word Grammar* (Blackwell, Oxford, 1984)
32. I. Mel'čuk, *Dependency Syntax: Theory and Practice* (State of New York University Press, Albany, 1988)
33. B. Bollobás, *Modern Graph Theory*. Graduate Texts in Mathematics (Springer, New York, 1998)
34. R. Ferrer i Cancho, R.V. Solé, in *Statistical Mechanics of Complex Networks*, *Lecture Notes in Physics*, vol. 625, ed. by R. Pastor-Satorras, J. Rubí, A. Díaz-Guilera (Springer, Berlin, 2003), pp. 114–125. DOI 10.1007/b12331
35. A.E. Goldberg, Trends in Cognitive Sciences **7**(5), 219 (2003). DOI 10.1016/S1364-6613(03)00080-9
36. C. Gómez-Rodríguez, D. Fernández-González, V.M.D. Bilbao, Computational Intelligence (2014). DOI 10.1111/coin.12027
37. M. Noy, Discrete Mathematics **180**, 301 (1998). DOI 10.1016/S0012-365X(97)00121-0
38. R. Ferrer-i-Cancho, http://arxiv.org/abs/1305.4561 (2013)
39. R. Ferrer-i-Cancho, Glottometrics **25**, 1 (2013)
40. D. Aldous, SIAM J. Disc. Math. **3**, 450 (1990). DOI 10.1137/0403039
41. A. Broder, in *Symp. Foundations of Computer Sci., IEEE* (New York, 1989), pp. 442–447
42. J. Moon, in *Canadian Math. Cong.* (1970)
43. H. Liu, Lingua **120**(6), 1567 (2010). DOI 10.1016/j.lingua.2009.10.001
44. N.J.A. Seoane, *Number of Trees with n Unlabeled Nodes* (2013). Available at http://oeis.org/A000055
45. R.V. Solé, Complexity **16**(1), 20 (2010). DOI 10.1002/cplx.20326
46. R. Ferrer-i-Cancho, N. Forns, A. Hernández-Fernández, G. Bel-Enguix, J. Baixeries, Complexity **18**(3), 11 (2013). DOI 10.1002/cplx.21429
47. H. Liu, Journal of Cognitive Science **9**, 159 (2008)
48. R. Ferrer-i-Cancho, http://arxiv.org/abs/1310.5884 (2013)
49. R. Ferrer-i-Cancho, Europhysics Letters **76**(6), 1228 (2006). DOI 10.1209/epl/i2006-10406-0
50. G. Morrill, O. Valentín, M. Fadda, in *Logic, Language, and Computation*, *Lecture Notes in Computer Science*, vol. 5422, ed. by P. Bosch, D. Gabelaia, J. Lang (Springer Berlin Heidelberg, 2009), pp. 272–286. DOI 10.1007/978-3-642-00665-4_22
51. R. Ferrer-i-Cancho, in *THE EVOLUTION OF LANGUAGE - Proceedings of the 10th International Conference (EVOLANG10)*, ed. by E.A. Cartmill, S. Roberts, H. Lyn, H. Cornish (Wiley, Vienna, Austria, 2014), pp. 66–73. DOI 10.1142/9789814603638_0007. Evolution of Language Conference (Evolang 2014), April 14-17
52. R. Ferrer-i-Cancho, Language Dynamics and Change **4**, in press (2015). URL `http://arxiv.org/abs/1309.1939`
53. F.R.K. Chung, Comp. & Maths. with Appls. **10**(1), 43 (1984). DOI 10.1016/0898-1221(84)90085-3
54. R. Ferrer-i-Cancho, H. Liu, Glottotheory **5**, 143 (2014). DOI 10.1515/glot-2014-0014

55. G. Morrill, Computational Linguistics **25**(3), 319 (2000). DOI 10.1162/089120100561728
56. J.A. Hawkins, in *Constituent Order in the Languages of Europe*, ed. by A. Siewierska (Mouton de Gruyter, Berlin, 1998)
57. R. Levy, E. Fedorenko, M. Breen, E. Gibson, Cognition **122**(1), 12 (2012). DOI 10.1016/j.cognition.2011.07.012
58. E. Gibson, E. Fedorenko, Trends in Cognitive Sciences **14**(6), 233 (2010). DOI 10.1016/j.tics.2010.03.005
59. P.W. Culicover, R. Jackendoff, Trends in Cognitive Sciences **14**(6), 234 (2010). DOI 10.1016/j.tics.2010.03.012
60. A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chatter, M. Christiansen, Trends in Cognitive Sciences **17**, 348 (2013). DOI 10.1016/j.tics.2013.04.010
61. R.A. Hochberg, M.F. Stallmann, Information Processing Letters **87**, 59 (2003). DOI 10.1016/S0020-0190(03)00261-8
62. H. Liu, Glottometrics **15**, 1 (2007)
63. M. Gell-Mann, M. Ruhlen, Proceedings of the National Academy of Sciences USA **108**(42), 17290 (2011). DOI 10.1073/pnas.1113716108
64. E. Gibson, in *Image, Language, Brain* (The MIT Press, Cambridge, MA, 2000), pp. 95–126
65. J.A. Hawkins, *A Performance Theory of Order and Constituency* (Cambridge University Press, New York, 1994)
66. M. Molloy, B. Reed, Random Structures and Algorithms **6**, 161 (1995). DOI 10.1002/rsa.3240060204
67. M. Molloy, B. Reed, Combinatorics, Probability and Computing **7**, 295 (1998). DOI 10.1017/S0963548398003526
68. M.E.J. Newman, S.H. Strogatz, D.J. Watts, Phys. Rev. E **64**, 026118 (2001). DOI 10.1103/PhysRevE.64.026118
69. E. Gibson, T. Warren, Syntax **7**, 55 (2004). DOI 10.1111/j.1368-0005.2004.00065.x
70. M.H. de Vries, K.M. Petersson, S. Geukes, P. Zwitserlood, M.H. Christiansen, Philosophical Transactions of the Royal Society B: Biological Sciences **367**(1598), 2065 (2012). DOI 10.1098/rstb.2011.0414
71. M. Bunge, *La ciencia. Su método y su filosofía* (Laetoli, Pamplona, 2013)
72. G.K. Zipf, *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Cambridge (MA), USA, 1949)
73. E. Bach, C. Brown, W. Marslen-Wilson, Language and Cognitive Processes **1**, 249 (1986). DOI 10.1080/01690968608404677
74. N. Evans, S.C. Levinson, Behavioral and Brain Sciences **32**, 429 (2009). DOI 10.1017/S0140525X0999094X
75. S. Piantadosi, E. Gibson, Cognitive Science **38**(4), 736 (2014). DOI 10.1111/cogs.12088