Memory limitations are hidden in grammar

Carlos Gómez-Rodríguez[1], Morten H. Christiansen[2,3], Ramon Ferrer-i-Cancho[4*]

[1] Universidade da Coruña, CITIC, FASTPARSE Lab, LyS Research Group, Depto. de Ciencias de la Computación y Tecnologías de la Información, A Coruña, Spain.
[2] Department of Psychology, Cornell University, Ithaca, NY, USA.
[3] Interacting Minds Centre and School of Communication and Culture, Nobelparken, Aarhus University, Denmark.
[4] Complexity and Quantitative Linguistics Lab, LARCA Research Group, Departament de Ciències de la Computació, Universitat Politècnica de Catalunya (UPC), Barcelona, Catalonia, Spain.
[*] Corresponding author's email: rferrericancho@cs.upc.edu

DOI: https://doi.org/10.53482/2022_52_397

ABSTRACT

The ability to produce and understand an unlimited number of different sentences is a hallmark of human language. Linguists have sought to define the essence of this generative capacity using formal grammars that describe the syntactic dependencies between constituents, independent of the computational limitations of the human brain. Here, we evaluate this independence assumption by sampling sentences uniformly from the space of possible syntactic structures. We find that the average dependency distance between syntactically related words, a proxy for memory limitations, is less than expected by chance in a collection of state-of-the-art classes of dependency grammars. Our findings indicate that memory limitations have permeated grammatical descriptions, suggesting that it may be impossible to build a parsimonious theory of human linguistic productivity independent of non-linguistic cognitive constraints.

Keywords: dependency syntax, dependency distance minimization, memory, grammar, network science

1 Introduction

An often celebrated aspect of human language is its capacity to produce an unbounded number of different sentences (Chomsky, 1965; Miller, 2000). For many centuries, the goal of linguistics has been to capture this capacity by a formal description—a grammar—consisting of a systematic set of rules and/or principles that determine which sentences are part of a given language and which are not (Bod, 2013). Over the years, these formal grammars have taken many forms but common to them all is the assumption that they capture the idealized linguistic competence of a native speaker/hearer, independent of any memory limitations or other non-linguistic cognitive constraints (Chomsky, 1965; Miller, 2000). These abstract formal descriptions have come to play a foundational role in the language sciences, from linguistics, psycholinguistics, and neurolinguistics (Hauser et al., 2002; Pinker, 2003) to computer science, engineering, and machine learning (Dyer et al., 2016; Gómez-Rodríguez et al., 2018; Klein and Manning, 2003). Despite evidence that processing difficulty underpins the unacceptability of certain sentences (Hawkins, 2004; Morrill, 2010), the cognitive independence assumption that is a defining feature of linguistic competence has not been examined in a systematic way using the tools of formal grammar. It is therefore unclear whether these supposedly idealized descriptions of language are free of non-linguistic cognitive constraints, such as memory limitations.

If the cognitive independence assumption should turn out not to hold, then it would have wide-spread theoretical and practical implications for our understanding of human linguistic productivity. It would require a reappraisal of key parts of linguistic theory that hitherto have posed formidable challenges for explanations of language processing, acquisition and evolution (Gold, 1967; Hauser et al., 2002; Pinker, 2003)—pointing to new ways of thinking about language that may simplify the problem space considerably by making it possible to explain apparently arbitrary aspects of linguistic structure in terms of general learning and processing biases (Christiansen and Chater, 2008; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). In terms of practical ramifications, engineers may benefit from building human cognitive limitations directly into their natural language processing systems, so as to better mimic human language skills and thereby improve performance. Here, we therefore evaluate the cognitive independence assumption using a state-of-the-art grammatical framework, dependency grammar (Nivre, 2005), to search for possible hidden memory constraints in these formal, idealized descriptions of natural language.

In dependency grammar, the syntactic structure of a sentence is defined by two components. First, a directed graph where vertices are words and arcs indicate syntactic dependencies between a head and its dependent. Such a graph has a root (a vertex that receives no edges) and edges are oriented away from the root (Figure 1). Second, the linear arrangement of the vertices of the graph (defined by the sequential order of the words in a sentence). Thus, syntactic dependency structures constitute a particular kind of spatial network where the graph is embedded in one dimension (Barthélemy, 2018), a correspondence that has led to the development of syntactic theory from a network theory standpoint (Gómez-Rodríguez and Ferrer-i-Cancho, 2017).

Dependency grammar is an important framework for various reasons. First, categorial grammar defines the syntactic structure of a sentence as dependency grammar (Morrill, 2010). Second, equivalences exist between certain formalisms of dependency grammar and constituency grammar (Gaifman, 1965; Kahane and Mazziotta, 2015). Third, there has been an evolution of minimalism towards dependency grammar (Osborne et al., 2011). Finally, dependency grammar has become a *de facto* standard in computational

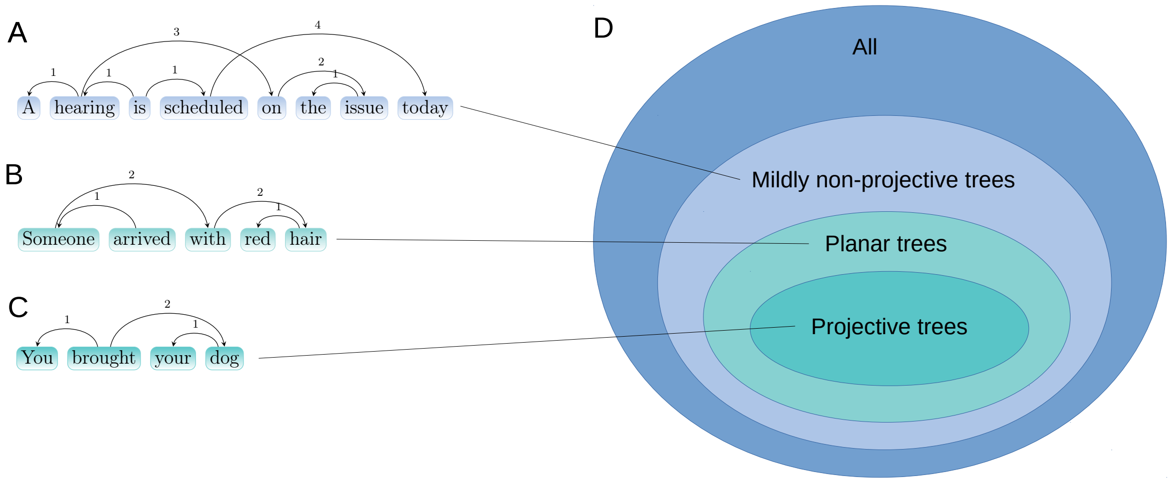

Figure 1: Examples of syntactic dependency structures. Arcs indicate syntactic dependencies from a head to its dependent and are labelled with the distance between them (distance is measured in words; consecutive words are at distance 1). n is the number of words of the sentence, D is the sum of dependency distances and $\langle d \rangle = D/(n-1)$ is the average dependency distance. A. A mildly non-projective tree from the classes $1EC$ and MH_4 (adapted from Nivre, 2009) where $n = 8$ and $\langle d \rangle = 13/7 \approx 1.85$. B. A planar but non-projective tree where $n = 5$ and $\langle d \rangle = 3/2$ (adapted from Groß and Osborne, 2009). C. A projective tree (adapted from Groß and Osborne, 2009) where $n = 4$ and $\langle d \rangle = 4/3$. D. A diagram of the superset relationships between projective, planar, mildly non-projective and unrestricted (all) syntactic dependency structures.

linguistics (Kübler et al., 2009).

To delimit the set of possible grammatical descriptions, various classes or sets of syntactic dependency structures have been proposed. These classes can be seen as filters on the possible linear arrangements of a given tree. Here, we consider four main classes. First, consider planar structures, where edges do not cross when drawn above the words of the sentence. The structure in Figure 1 B-C are planar while that of Figure 1 A is not. Second, we have projective structures, the most well-known class. A dependency tree is projective if, and only if, it is planar and its root is not covered by any dependency (Figure 1 C). Third, there are mildly non-projective structures, comprising the union of planar structures and additional structures with further (but slight) deviations from projectivity, e.g., by having a low number of edge crossings (Figure 1 A). Finally, the class of all structures, that has no constraints on the possible structures.

Figure 1 D shows the inclusion relationships among these classes. However, the whole picture, encompassing state-of-the-art classes is more complex. Mildly non-projective structures are not actually a class but a family of classes. We have selected three representative classes: MH_k, WG_1 and $1EC$ structures, that are supersets of projective structures but whose definition is more complex (see Methods).

Here we validate the assumption of independence between grammatical constraints and cognitive limitations in these classes of grammar using the distance between syntactically related words in a dependency tree as a proxy for memory constraints (Liu et al., 2017; Temperley and Gildea, 2018). Such a distance is defined as the number of intermediate words plus one. Thus, if the linked words are consecutive they are at distance 1, if they are separated by an intermediate word they are at distance two, and so on, as shown in Figure 1. Dependency distance minimization is a pressure to reduce the distance between syntactically related words that is supported statistically by large-scale analyses of syntactic dependency structures in languages (Ferrer-i-Cancho et al., 2022; Futrell et al., 2020; Futrell et al., 2015; Jing et al., 2021; Liu, 2008). As such, dependency distance minimization is a type of memory constraint, believed to result from pressure against decay of activation or interference during the processing of sentences (Liu et al., 2017; Temperley and Gildea, 2018). Dependency distances tax memory and cognition in general. Dependency distances reduce in case of cognitive impairment (Aronsson et al., 2021; Roark et al., 2011). There is an association between the level of cognitive impairment and dependency distance: as the severity of the impairment increases, dependency distances tend to be reduced (Aronsson et al., 2021). Moreover, an association between the level of competence of L2 learners and dependency distance has also been found: as learners of a second language become more competent in the new language, dependency distances increase (Ouyang and Jiang, 2018; Yuan et al., 2021).

The article is written so that reading the next section, *Materials and methods* (Section 2) is not essential to understand the *Results* section (Section 3). Therefore, it is up to reader to decide whether to proceed with Section 2 or to skip to Section 3, reading Section 2 later on.

2 Material and Methodology

2.1 Control for Sentence Length

In our study, we do not investigate the average dependency distance over a whole ensemble of dependency structures but instead we condition on sentence length (Ferrer-i-Cancho and Liu, 2014; Futrell et al., 2015). Then for a given n, we calculate $\langle d \rangle_{AS}$, the average dependency length for an ensemble of artificial syntactic dependency structures (AS), and also $\langle d \rangle_{RS}$, the average dependency length for an ensemble of attested syntactic dependency structures (RS). By doing that, we are controlling for sentence length, getting rid of the possible influence of the distribution of sentence length in the calculation of $\langle d \rangle_{RS}$ or $\langle d \rangle_{AS}$ (Ferrer-i-Cancho and Liu, 2014).

2.2 Attested Syntactic Dependency Structures

We estimated the average dependency distances in attested sentences using collections of syntactic dependency treebanks from different languages. A syntactic dependency treebank is a database of sentences and their syntactic dependency trees.

To provide results on a wide range of languages while controlling for the effects of different syntactic annotation theories, we use two collections of treebanks:

- Universal Dependencies (UD), version 2.4 (Nivre et al., 2019). This is the largest available collection of syntactic dependency treebanks, featuring 146 treebanks from 83 distinct languages. All of these treebanks are annotated following the common Universal Dependencies annotation criteria, which are a variant of the Stanford Dependencies for English (de Marneffe and Manning, 2008), based on lexical-functional grammar (Bresnan, 2000), adapting them to be able to represent syntactic phenomena in diverse languages under a common framework. This collection of treebanks can be freely downloaded[1] and is available under free licenses.

- HamleDT 2.0 (Rosa et al., 2014). This collection is smaller than UD, featuring 30 languages, all of which (except for one: Bengali) are also available in UD, often with overlapping source material. Thus, using this collection does not meaningfully extend the diversity of languages covered beyond using only UD. However, the interest of HamleDT 2.0 lies in that each of the 30 treebanks is annotated with not one, but two different sets of annotation criteria: Universal Stanford dependencies (de Marneffe et al., 2014) and Prague Dependencies (Hajič et al., 2006). We abbreviate these two

[1] https://universaldependencies.org/. Last accessed 17 February 2022.

subsets of the HamleDT 2.0 collection as "Stanford" and "Prague", respectively. While Universal Stanford dependencies are closely related to UD, Prague dependencies provide a significantly different view of syntax, as they are based on the functional generative description (Sgall, 1969) of the Praguian linguistic tradition (Hajicova, 1995), which differs from Stanford dependencies in substantial ways, like the annotation of conjunctions or adpositions (Passarotti, 2016). Thus, using this version of HamleDT[2] makes our analysis more robust, as we can draw conclusions without being tied to a single linguistic tradition. The HamleDT 2.0 treebanks are available online.[3] While not all of the treebanks are made fully available to the public under free licenses, to reproduce our analysis it is sufficient to use a stripped version where the words have been removed from the sentences for licensing reasons, but the bare trees are available. This version is distributed freely.[4]

A preprocessed file with the minimal information needed to reproduce our measurements on attested syntactic structures (Figure 6 A) is available. [5]

To preprocess the treebanks for our analysis, we removed punctuation, following common practice in statistical research of dependency structures (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). We also removed tree nodes that do not correspond to actual words, such as the null elements in the Bengali, Hindi and Telugu HamleDT corpora and the empty nodes in several UD treebanks. To ensure that the dependency structures are still valid trees after these removals, we reattached nodes whose head has been deleted as dependents of their nearest non-deleted ancestor. Finally, in our analysis we disregarded syntactic trees with less than three nodes, as their statistical properties are trivial and provide no useful information (a single-node dependency tree has no dependencies at all, and a 2-node tree always has a single dependency of distance 1). Tables 1 and 2 summarize the languages in each collection of treebanks.

2.3 Artificial Syntactic Dependency Structures

Apart from the attested trees, we used a collection of over 16 billion randomly-generated trees. For values of n (the length or number of nodes) from 3 to $n^* = 10$, we exhaustively obtained all possible trees. The number of possible dependency trees for a given length n is given by n^{n-1}, ranging from 9 possible trees for $n = 3$ to 10^9 for $n = n^*$. From $n > n^*$ onwards, the number of trees grows too large to be manageable, so we resort to uniformly random sampling of 10^9 trees for $n^* < n \leq 25$. For each tree

[2]While there is a later version (HamleDT 3.0), it abandoned the dual annotation and adopted Universal Dependencies instead, thus making it less useful for our purposes.

[3]https://ufal.mff.cuni.cz/hamledt/hamledt-treebanks-20. Last accessed 17 February 2022.

[4]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-9551-4?show=full. Last accessed 17 February 2022.

[5]https://doi.org/10.7910/DVN/XHRIYX

Table 1: The languages in the UD collection grouped by family. The counts attached to the collection name indicate the number of different families and the number of different languages. The counts attached to family names indicate the number of different languages.

Collection	Family	Languages
UD (19, 83)	Afro-Asiatic (7)	Akkadian, Amharic, Arabic, Assyrian, Coptic, Hebrew, Maltese
	Turkik (3)	Kazakh, Turkish, Uyghur
	Austro-Asiatic (1)	Vietnamese
	Austronesian (2)	Indonesian, Tagalog
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (46)	Afrikaans, Ancient Greek, Armenian, Belarusian, Breton, Bulgarian, Catalan, Croatian, Czech, Danish, Dutch, English, Faroese, French, Galician, German, Gothic, Greek, Hindi, Hindi-English, Irish, Italian, Kurmanji, Latin, Latvian, Lithuanian, Marathi, Norwegian, Old Church Slavonic, Old French, Old Russian, Persian, Polish, Portuguese, Romanian, Russian, Sanskrit, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Upper Sorbian, Urdu, Welsh
	Japanese (1)	Japanese
	Korean (1)	Korean
	Mande (1)	Bambara
	Mongolic (1)	Buryat
	Niger-Congo (2)	Wolof, Yoruba
	Other (1)	Naija
	Pama-Nyungan (1)	Warlpiri
	Sign Language (1)	Swedish Sign Language
	Sino-Tibetan (3)	Cantonese, Chinese, Classical Chinese
	Tai-Kadai (1)	Thai
	Tupian (1)	Mbya Guarani
	Uralic (7)	Erzya, Estonian, Finnish, Hungarian, Karelian, Komi Zyrian, North Sami

Table 2: The languages in the HamleDT collections (Stanford and Prague) grouped by family. The counts attached to the collection names indicate the number of different families and the number of different languages. The counts attached to family names indicate the number of different languages.

Collection	Family	Languages
Stanford (7, 30)	Afro-Asiatic (1)	Arabic
	Turkik (1)	Turkish
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (21)	Ancient Greek, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, English, German, Greek, Hindi, Italian, Latin, Persian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish
	Japanese (1)	Japanese
	Uralic (3)	Estonian, Finnish, Hungarian
Prague (7, 30)	Afro-Asiatic (1)	Arabic
	Turkik (1)	Turkish
	Basque (1)	Basque
	Dravidian (2)	Tamil, Telugu
	Indo-European (21)	Ancient Greek, Bengali, Bulgarian, Catalan, Czech, Danish, Dutch, English, German, Greek, Hindi, Italian, Latin, Persian, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish
	Japanese (1)	Japanese
	Uralic (3)	Estonian, Finnish, Hungarian

in the collection, the classes it belongs to are indicated in the dataset[6].

The reason why we do not go beyond length 25 is that, for larger lengths, trees that belong to our classes under analysis are very scarce (Figure 2 A). For example, even sampling 10^9 random trees for each length, no projective trees are found for $n > 18$. The same can be said of planar trees for $n > 19$, $1EC$ trees for $n > 22$, MH_4 trees for $n > 23$, and WG_1 trees for $n > 24$. For the MH_5 class, some trees can still be found in the sample for length 25, but only 69 out of 10^9 belong to the class. Due to undersampling, the plot on artificial structures in the results section only shows points represented by at least 30 structures for $n > n^*$. 30 is considered a rule of thumb for the minimum sample size that is needed to estimate the mean of random variables that follow short tailed distributions (Hogg and Tanis, 1997). Figure 2 B shows average dependency distances not excluding any point.

For $n \leq n^*$, the ensemble of AS used to calculate $\langle d \rangle_{AS}$ contains all possible syntactic dependency structures for all classes. For $n > n^*$, it contains a random sample of them. Within a given ensemble, each structure is generated from a labelled directed tree whose vertex labels are interpreted as vertex positions in the linear arrangement. The values of $\langle d \rangle_{AS}$ for each class are exact (the mean over all possible syntactic dependency structures) for $n \leq n^*$ and random sampling estimates for $n > n^*$. A detailed explanation follows.

For a given n, an ensemble of syntactic dependency structures is generated with a procedure that is a generalization of the procedure used to generate random structures formed by an undirected tree and a linear arrangement (Esteban et al., 2016). The procedure has two versions: the exhaustive version, that was used for $n \leq n^*$, and the random sampling version, that was used for $n > n^*$. The exhaustive version consists of

1. Generating all the $T(n)$ labelled (undirected) trees of n vertices using Prüfer codes (Prüfer, 1918). It is known that $T(n) = n^{n-2}$ (Cayley, 1889).

2. Converting each of these random trees into labelled directed trees (i.e., dependency trees) by rooting it in all possible ways. A rooting consists in choosing one node of the tree as the root, and making all edges point away from the root via a depth-first traversal. This produces $nT(n) = n^{n-1}$ syntactic dependency structures.

3. Producing a syntactic dependency structure from every directed tree using vertex labels (integers from 1 to n) as vertex positions in a linear arrangement (Esteban et al., 2016).

[6]The trees are freely available from https://doi.org/10.7910/DVN/XHRIYX

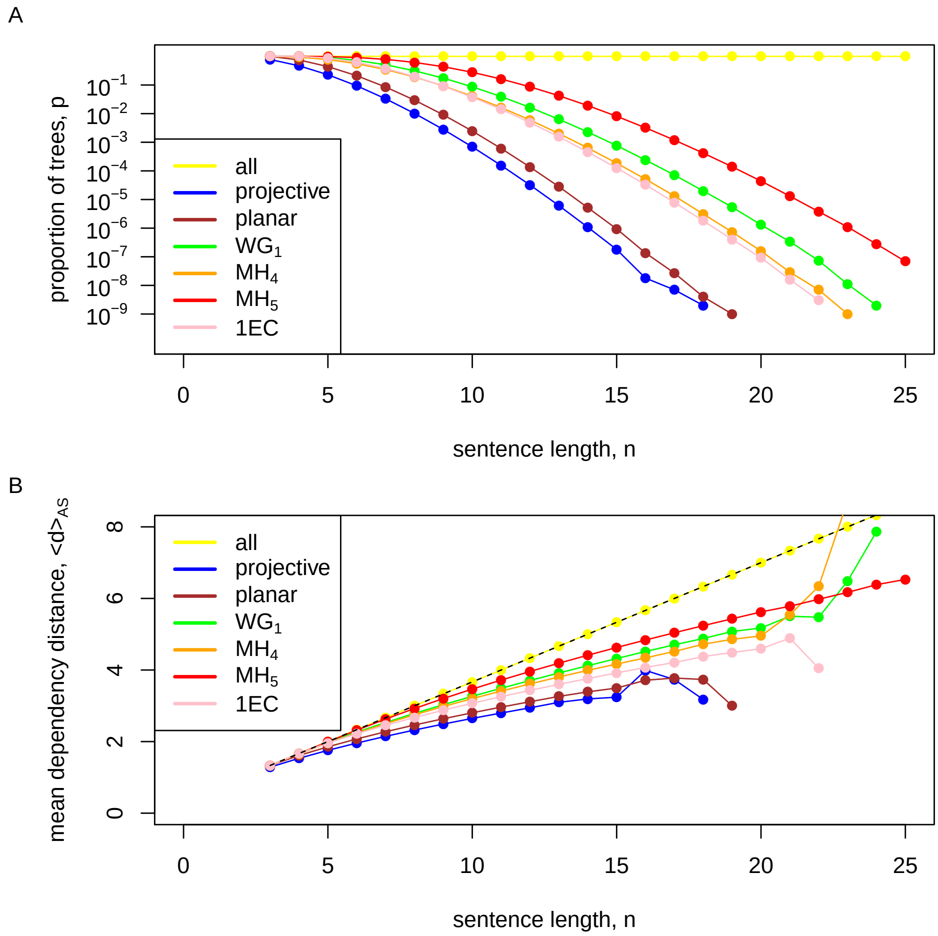

Figure 2: Undersampling in artificial syntactic dependency structures (AS). A. p, the proportion of artificial structures of a certain class in the sample. B. The average dependency length (in words), $\langle d \rangle_{AS}$, as a function of n, the sentence length (in words). For reference, the base line defined by a random linear arrangement of the words of the sentence, $\langle d \rangle_{rla}$ (Eq. 3) is also shown (dashed line).

4. Discarding the trees that do not belong to the target class.

The random sampling version consists of

1. Generating S uniformly random labelled (undirected) trees of n vertices, via uniformly random Prüfer codes (Prüfer, 1918).
2. Converting these uniformly random labelled trees to uniformly random labelled directed trees (i.e., dependency trees) by randomly choosing one node of each tree as the root, and making all edges point away from the root via a depth-first traversal. This produces S syntactic dependency structures.
3. Same as exhaustive version.
4. Same as exhaustive version.

Note that Step 2 warrants that labelled directed trees in the ensemble are uniformly random: if we call K_n the probability of generating each undirected tree of n vertices with a random Prüfer code, we can observe that each possible directed tree corresponds to exactly one undirected tree (obtained by ignoring arc directions), and each undirected tree corresponds to exactly n distinct directed trees (resulting from picking each of its n nodes as the root). Thus, the method of generating a random Prüfer code and then choosing a root generates each possible directed tree with a uniform probability K_n/n (as the probability of choosing the underlying undirected tree is K_n, and the probability of choosing the relevant root is $1/n$).

After each procedure, the average dependency length $\langle d \rangle$ for a given n and a given class is calculated. While the exhaustive procedure allows one to calculate the true average dependency length over a certain class, the random sampling algorithm only allows one to estimate the true average. Put differently, the exhaustive procedure allows one to calculate exactly the expected dependency length in a class assuming that all labelled directed trees are equally likely whereas the random sampling procedure only allows one to obtain an approximation.

We explore all values of n within the interval $[n_{min}, n_{max}]$ with $n_{min} = 3$ and $n_{max} = 25$ and $n^* = 10$ and $S = 10^9$. The total number of syntactic dependency structures generated for our study is

$$U = (n_{max} - n^*)S + \sum_{n=n_{min}}^{n^*} nT(n) = (n_{max} - n^*)S \sum_{n=n_{min}}^{n^*} n^{n-1}.$$

Applying the parameters above, one obtains

$$U \approx 1.6 \cdot 10^{10} \tag{1}$$

2.4 The Random Baseline

Although the random baseline

(2) $$\langle d \rangle_{rla} = (n+1)/3$$

follows from Jaynes' maximum entropy principle in the absence of any constraint (Kesavan, 2009), it may be objected that our baseline is too unconstrained from a linguistic perspective. In previous research, random baselines that assume projectivity or consistent branching, whereby languages tend to grow parse trees either to the right (as in English) or to the left (as in Japanese), have been considered (Futrell et al., 2015; Gildea and Temperley, 2010; Liu, 2008). However, it has been argued that these linguistic constraints could be a reflection of memory limitations (Christiansen and Chater, 1999; Ferrer-i-Cancho and Gómez-Rodríguez, 2016b). Therefore, incorporating these linguistic constraints into the baseline for evaluating dependency distances would not provide an adequate test of the cognitive independence assumption because they could mask the effect of dependency distance minimization (DDm). Consistently, the planarity assumptions reduces the statistical power of a test of DDm (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). In addition, these additional constraints compromise the parsimony of a general theory of language for neglecting the predictive power of DDm (Ferrer-i-Cancho and Gómez-Rodríguez, 2016b).

A priori, $\langle d \rangle_{AS}$ could be below the random baseline as it occurs typically in human languages (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho and Liu, 2014) but it could also be above. As for the latter situation, empirical research in short sentences has shown that there are languages where dependency lengths are larger than expected by chance (Ferrer-i-Cancho and Gómez-Rodríguez, 2021). In addition, there exist syntactic dependency structures where $\langle d \rangle > \langle d \rangle_{rla}$ from a network theoretical standpoint. For instance, among planar syntactic structures, the maximum average dependency distance is $\langle d \rangle_{max} = n/2$ (Ferrer-i-Cancho, 2013).

$\langle d \rangle_{AS}$ never exceeds $\langle d \rangle_{rla}$ and it deviates from $\langle d \rangle_{rla}$ when $n = 3$ for projective trees, $n = 4$ for planar trees and MH_4 and $n = 5$ for $1EC$, MH_5 and WG_1. For the class of all syntactic dependency structures (Figure 1 D), we find that $\langle d \rangle_{AS}$ matches Eq. 2 as expected from previous research (Esteban et al., 2016).

2.5 The Classes of Dependency Structures

Planar trees: A dependency tree is said to be *planar* (or *noncrossing*) if its dependency arcs do not cross when drawn above the words. Planar trees have been used in syntactic parsing algorithms (Gómez-Rodríguez and Nivre, 2010), and their generalization to noncrossing graphs has been widely studied both for its formal properties (Yli-Jyrä and Gómez-Rodríguez, 2017) and for parsing (Kuhlmann and Jonsson, 2015).

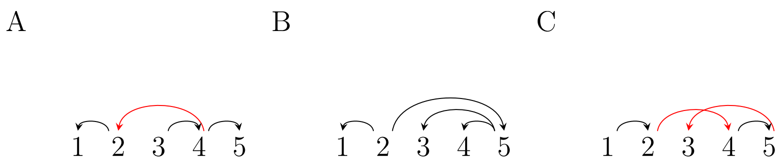

Figure 3: Planarity and projectivity. A. A tree that is planar (dependencies do not cross) but not projective (the root node, 3, is covered by the dependency in red). B. A tree that is planar and projective. C. A tree that is not planar (the dependencies in red cross), and thus not projective.

Projective trees: A dependency tree is said to be projective if it is planar and its root is not covered by any dependency (Figure 3). Projectivity facilitates the design of simple and efficient parsers (Nivre, 2003, 2004), whereas extending them to support non-projective trees increases their computational cost (Covington, 2001; Nivre, 2009). For this reason, and because treebanks of some languages (like English or Japanese) have traditionally had few or no non-projective analyses, many practical implementations of parsers assume projectivity (Chen and Manning, 2014; Dyer et al., 2015).

However, non-projective parsing is needed to deal with sentences exhibiting non-projective phenomena such as extraposition, scrambling or topicalization. Non-projectivity is particularly common in flexible word order languages, but generally present in a wide range of languages. However, non-projectivity in natural languages tends to be *mild* in the sense that the actually occurring non-projective trees are very close to projective trees, as they have much fewer crossing dependencies than would be expected by chance (Ferrer-i-Cancho et al., 2018).

For this reason, there has been research interest in finding a restriction that would be a better fit for the phenomena observed in human languages. From a linguistic standpoint, the goal is to describe the syntax of human language better than with the overly restrictive projective trees or the arguably excessive permissiveness of admitting any tree without restriction, disregarding the observed scarcity of crossing dependencies. From an engineering standpoint, the goal is to strike a balance between the efficiency provided by more restrictive parsers with a smaller search space and the coverage of the non-projective phenomena that can be found in attested sentences. In this line, various sets of dependency structures that have been proposed are supersets of projective trees allowing only a limited degree of non-projectivity. These sets are called mildly non-projective classes of dependency trees (Kuhlmann and Nivre, 2006).

Here, we focus on three of the best known such sets, which have interesting formal properties and/or have been shown to be practical for parsing due to providing a good efficiency-coverage trade-off. We briefly outline them here, and refer the reader to Gómez-Rodríguez, 2016 for detailed definitions and coverage

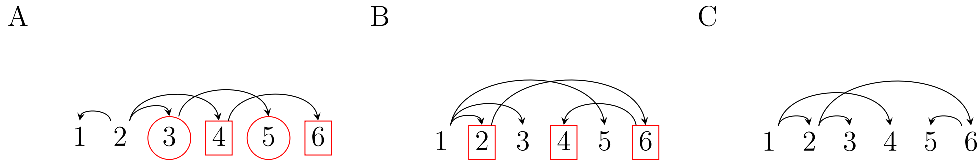

Figure 4: Well-nestedness and gap degree. A. An ill-nested tree (the yields of node 3—circled—and node 4–squared—form an interleaving pattern). B. A tree with gap degree 2 (the yield of node 2, squared, has two discontinuities, at nodes 3 and 5). C. A tree that is well-nested and has gap degree 1, and thus is in WG_1.

statistics of these and other mildly non-projective classes of trees.

Well-nested trees with Gap degree 1 (WG_1): A dependency tree is well-nested (Bodirsky et al., 2005) if it does not contain two nodes with disjoint, interleaving yields. Given two disjoint yields $a_1 \dots a_p$ and $b_1 \dots b_q$, they are said to interleave if there exist i, j, k, l such that $a_i < b_j < a_k < b_l$. On the other hand, the gap degree of a tree is the maximum number of discontinuities present in the yield of a node, i.e., a dependency tree has gap degree 1 if every yield is either a contiguous substring, or the union of two contiguous substrings of the input sentence. Figure 4 provides graphical examples of these properties. WG_1 trees have drawn interest mainly from the formal standpoint, for their connections to constituency grammar (Kuhlmann, 2010), but they also have been investigated in dependency parsing (Corro et al., 2016; Gómez-Rodríguez et al., 2011; Gómez-Rodríguez et al., 2009).

Multi-Headed with at most k heads per item (MH_k): Given $k \geq 3$, the set of MH_k trees contains the trees that can be parsed by an algorithm called MH_k (Gómez-Rodríguez et al., 2011). k is a parameter of the class, such that for $k = 3$ the class coincides with projective trees, but for $k > 3$ it covers increasingly larger sets of non-projective structures (but the parser becomes slower). A recent neural implementation of the MH_4 parser has obtained competitive accuracy on UD (Gómez-Rodríguez et al., 2018). For $k > 4$, the MH_k sets have been shown to be Pareto optimal (among known mildly non-projective classes) in terms of balance between efficiency and practical coverage (Gómez-Rodríguez, 2016). In this paper, we will consider the MH_4 and MH_5 sets.

1-Endpoint-Crossing trees ($1EC$): A dependency tree has the property of being 1-Endpoint-Crossing if, given a dependency, all other dependencies crossing it are incident to a common node (Pitler et al., 2013). This property is illustrated in Figure 5. 1EC trees were the first mildly non-projective class of dependency trees to have a practical exact-inference parser (Pitler, 2014), which was reimplemented with a neural architecture in (Gómez-Rodríguez et al., 2018). They are also in the Pareto frontier with respect to coverage and efficiency, according to Gómez-Rodríguez, 2016.

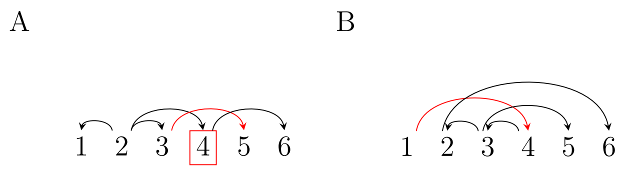

Figure 5: 1-Endpoint-Crossing property. A. A 1-Endpoint-Crossing tree (given any dependency, dependencies crossing it are incident to a common node—for example, here the dependencies crossing the one marked in red are incident to node 4). B. A tree that is not 1-Endpoint-Crossing. The dependency arc in red has two crossing dependencies which are not incident to any common node.

3 Results

3.1 Short Dependency Distances in Attested Structures Revisited

Assuming that all the linear arrangements are equally likely, $\langle d \rangle$, the average of dependency distances in a sentence of n words, is expected to be (Ferrer-i-Cancho, 2004)

(3) $$\langle d \rangle_{rla} = (n+1)/3.$$

Figure 6 A shows that $\langle d \rangle_{RS}$, the average dependency distance in attested syntactic dependency structures (RS), is below the random baseline defined by $\langle d \rangle_{rla}$ (see Methods for a justification of this baseline). This is in line with previous statistical analyses (Ferrer-i-Cancho, 2004; Futrell et al., 2015; Liu, 2008; Park and Levy, 2009) (see Liu et al., 2017; Temperley and Gildea, 2018 for a broader review of previous work) and the expected influence of performance constraints on attested trees.

The fact that $\langle d \rangle_{RS}$ is below 4 has been interpreted as a sign that dependency lengths are constrained by working memory limitations (Liu, 2008). For this reason, we test whether memory effects have permeated the classes of grammar by determining if $\langle d \rangle_{AS}$, the average dependency distance in a collection of artificial syntactic dependency structures (AS) from a certain class, is also below $\langle d \rangle_{rla}$ (Eq. 3). The purpose of Figure 6 A is merely to provide the reader with a baseline derived from attested dependency structures in natural language as a backdrop for the main contribution of the article, which is based on artificial syntactic dependency structures.

3.2 Short Dependency Distances in Artificial Structures

For a given n, we generate an ensemble of artificial syntactic dependency structures by exhaustive sampling for $n \leq n^* = 10$ and random sampling for $n > n^*$ (Methods). These artificial syntactic dependency trees are only constrained by the definition of the different classes. They are thus free

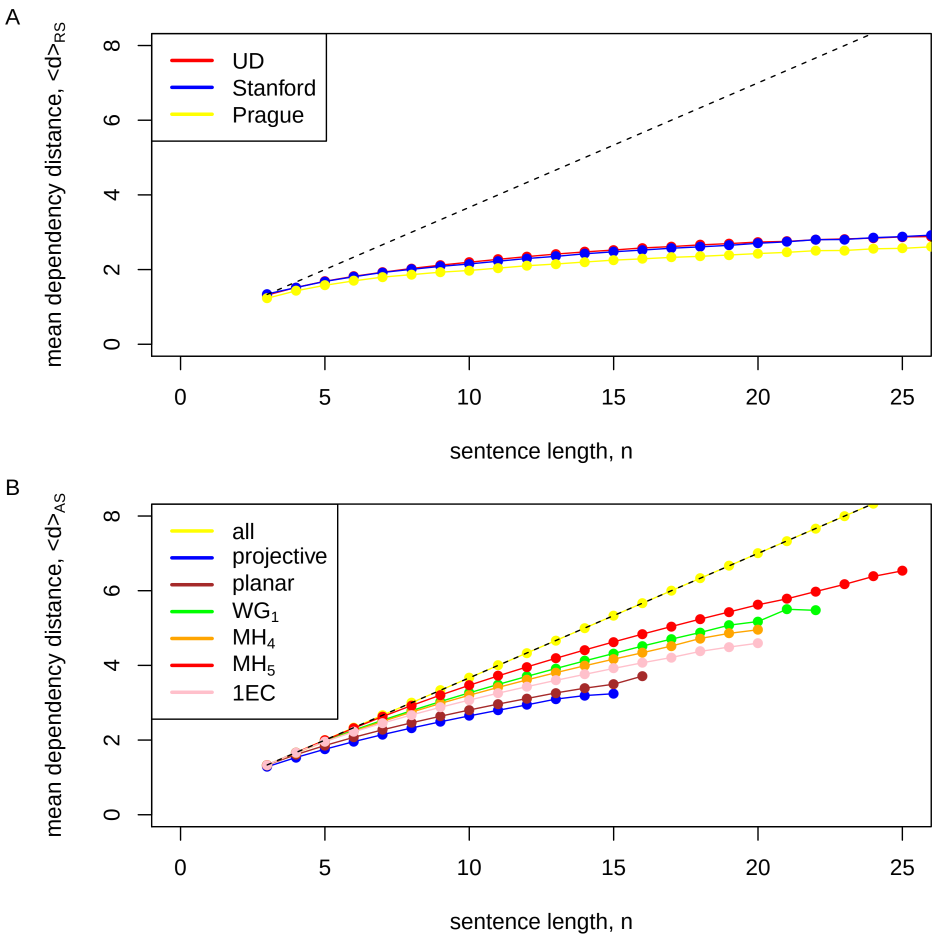

Figure 6: The average dependency length (in words), $\langle d \rangle$, as a function of n, the sentence length (in words). For reference, the baseline defined by a random linear arrangement of the words of the sentence, $\langle d \rangle_{rla}$ (Eq. 3) is also shown (dashed line). A. Attested syntactic dependency trees (RS) following three different annotation criteria: UD, Prague and Stanford dependencies. B. Artificial syntactic dependency structures (AS) belonging to different classes of grammars. Due to undersampling, only points represented by at least 30 structures are shown for $n > n^*$.

from any memory constraint other than the ones the different classes of grammars may, perhaps, impose indirectly. Still, these artificial syntactic structures have dependency lengths that are below the chance level (Figure 6 B), indicating that memory constraints are hidden in the definition of these classes. Interestingly $\langle d \rangle_{AS}$ is below chance for sufficiently large n in all classes of grammars although $\langle d \rangle_{AS}$ could be above $\langle d \rangle_{rla}$ (Eq. 3) in principle (see Methods). In general, the largest reduction of $\langle d \rangle_{AS}$ with respect to the random baseline is achieved by the projective class, followed by the planar class.

It is worth noting that a reduction of $\langle d \rangle_{AS}$ with respect to our random baseline has been observed for the projective class in past work, but with some important caveats: Liu, 2008 did not control for sentence length as in Figure 6 B; and whereas Park and Levy, 2009 did implement this control and considered another class of marginal interest (2-component structures) in addition to projective trees, their use of attested dependency trees instead of artificial control trees suggests that memory limitations might have influenced the results.

4 Discussion

The reduction of $\langle d \rangle$ with respect to the random baseline in artificial trees from a wide range of state-of-the art classes is consistent with the hypothesis that the scarcity of crossing dependencies is a side-effect of pressure to reduce the distance between syntactically related words (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). The smaller reduction of dependency distances with respect to the random baseline in artificial dependency structures can be explained by the fact that the curves in Figure 6 B derive from uniform sampling of the space of all possible trees. In contrast, real speakers may not only choose linear arrangements that reduce dependency distance, but also sample the space of possible structures with a bias towards structures that facilitate that such reduction or that satisfy other cognitive constraints (Ferrer-i-Cancho and Gómez-Rodríguez, 2021).

Our findings complete our understanding of the relationship between projectivity or mildly non-projectivity and dependency distance minimization. It has been shown that such minimization leads to a number of edge crossings that is practically zero (Ferrer-i-Cancho, 2006), and to not covering the root, one of the conditions for projectivity, in addition to planarity (Ferrer-i-Cancho, 2008). Here, we have demonstrated a complementary effect, i.e., that dependency distance reduces below chance when edge crossings are minimized (planarity) or projectivity is imposed. Whereas a recent study of similar classes of grammars suggested that crossing dependencies are constrained by either grammar or cognitive pressures rather than occurring naturally at some rate (Yadav et al., 2019), our findings strongly demonstrate that it is not grammar but rather non-linguistic cognitive constraints, that limit the occurrence of crossing dependencies in languages. Since we released the first version of this article in August 2019, https://arxiv.org/abs/1908.06629, other researchers have confirmed that dependency distance minimization contributes significantly to the emergence of formal constraints on crossing dependencies (Yadav et al., 2021, 2022). Yadav et al., 2021 have also confirmed the findings of previous research indicating that the effect of dependency distances alone leads to overestimate the actual number of crossing dependencies (Gómez-Rodríguez and Ferrer-i-Cancho, 2017); a critical point is that Gómez-Rodríguez and Ferrer-i-Cancho (2017) use a normalized score leading to the conclusion that such overestimation implies a small relative error.

We sampled about 16 billion syntactic dependency structures, that differed in length and syntactic complexity, to determine whether linguistic grammars are free of non-linguistic cognitive constraints, as is typically assumed. Strikingly, while previous work on natural languages has shown that dependency lengths are considerably below what would be expected by a random baseline without memory constraints (Ferrer-i-Cancho, 2004; Ferrer-i-Cancho and Liu, 2014; Liu, 2008; Park and Levy, 2009), we still observe a drop in dependency lengths for randomly generated, mildly non-projective structures that supposedly abstract away from cognitive limitations. Our interpretation of these results is that memory constraints, in the form of dependency distance minimization, have become inherent to formal linguistic grammars. We have demonstrated that distinct formal classes of mild non-projectivity manifest the sort of burden of dependency distances for memory and cognition that is observed in psychological experiments (Liu et al., 2017, Section 2) and that has been observed to become more marked in case of cognitive impairment (Aronsson et al., 2021) or second language learning (Ouyang and Jiang, 2018; Yuan et al., 2021).

It may be objected that our argument that memory limitations have permeated grammars is based on artificially generated syntactic structures instead of real ones. However, it is all but impossible to study real dependency structures without possible contamination from linguistic or non-linguistic cognitive constraints other than the formal mild non-projectivity classes. For that reason, here and in previous research (Ferrer-i-Cancho, 2014), we have focused on artificially generated syntactic structures. Notice this research is part of a larger research program were we have already used real syntactic dependency structures, but minimizing assumptions to argue that the scarcity of crossing dependencies can be explained to a large extent by dependency distance minimization (Gómez-Rodríguez and Ferrer-i-Cancho, 2017). Nonetheless, further research is needed with real syntactic dependency structures and the current study is a key, necessary step in this direction.

It may also be objected that our conclusions are limited by the sample of classes that we have considered and that we cannot exclude the possibility that, in the future, researchers might adopt a new class of mildly non-projective structures whose dependency distances cannot be distinguished from the random baseline. However, we believe that this is very unlikely for the following reasons: (1) our current sample of classes is representative of the state of the art (Gómez-Rodríguez, 2016), and spans classes that originated with different goals and motivations (from purely theoretical to parsing efficiency), with all sharing the drop in dependency lengths, (2) while one could conceivably engineer a class to have lengths in line with the baseline while still having high coverage of linguistic phenomena, this would mean forwarding more responsibility for dependency distance reduction to other parts of the linguistic theory in order to warrant that dependency distances are reduced to a realistic degree (Figure 6) and hence would preclude a parsimonious approach to language, and (3) given the positive correlation between crossings and dependency lengths (Alemany-Puig, 2019; Ferrer-i-Cancho and Gómez-Rodríguez, 2016a), such a class would be likely to have many dependency crossings, so it would be, at the least, questionable to call it mildly non-projective.

Beyond upending longheld assumptions about the nature of human linguistic productivity, our findings also have key implications for debates on how children learn language, how language evolved, and how computers might best master language. Whereas a common assumption in the acquisition literature is that children come to the task of language learning with built-in linguistic constraints on what they learn (Gold, 1967; Pinker, 2003), our results suggest that language-specific constraints may not be needed and instead be replaced by general cognitive constraints (Tomasello, 2005). The strong effects of memory on dependency distance minimization provide further support for the notion that language evolved through processes of cultural evolution shaped by the human brain (Christiansen and Chater, 2008), rather than the biological evolution of language-specific constraints (Pinker, 2003). Finally, our results raise the intriguing possibility that if we want to develop computer systems that target human linguistic ability in the context of human-computer interaction, we may paradoxically need to constraint the power of such systems to be in line with human cognitive limitations, rather than giving them the super-human computational capacity of AlphaGo. Memory limitations in the form of dependency minimization have already been applied to machine learning methods, but imposing planarity as if planarity and memory limitations were unrelated constraints (Eisner and Smith, 2010; Smith and Eisner, 2006, for instance). This suggests that considering planarity and other formal constraints as the effect of dependency minimization could boost machine learning methods

Our study was conducted using the framework of dependency grammar, but because of the close relationship between this framework and other ways of characterizing the human unbounded capacity to produce different sentences (Chomsky, 1965; Miller, 2000), such as categorial grammar (Morrill, 2010), phrase structure grammar (Gaifman, 1965; Kahane and Mazziotta, 2015), and minimalist grammar (Osborne et al., 2011), our results suggest that any parsimonious grammatical framework will incorporate memory constraints. Notice that, as a result of our study, we cannot refute the cognitive independence assumption. Our point is that the independence assumption leads to a less parsimonious theory of syntax. We are simply invoking Occam's razor so that formal constraints and the cognitive burden of dependency distances are not treated as separate entities. Moreover, given that dependency grammars constitute a special case of a graph that is embedded in one dimension, the physics toolbox associated with statistical mechanics and network theory may be applied to provide further insight into the nature of human linguistic productivity (Barthélemy, 2018; Gómez-Rodríguez and Ferrer-i-Cancho, 2017). However, these future explorations notwithstanding, our current findings show that memory limitations have permeated current linguistic conceptions of grammar, suggesting that it may not be possible to adequately capture our unbounded capacity for language, at least in the context of a parsimonious theory compatible with the idea of mild non-projectivity, without incorporating non-linguistic cognitive constraints into the grammar formalism.

Acknowledgments

This article is dedicated to the memory of G. Altmann, 1931-2020 (Köhler et al., 2021). We are grateful to L. Alemany-Puig, A. Hernandez-Fernandez and M. Vitevitch for helpful commments. CGR has received funding from the European Research Council (ERC), under the European Union's Horizon 2020 research and innovation programme (FASTPARSE, grant agreement No 714150), from ERDF/MICINN-AEI (ANSWER-ASAP, TIN2017-85160-C2-1-R; SCANNER-UDC, PID2020-113230RB-C21), from Xunta de Galicia (ED431C 2020/11), and from Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (ERDF - Galicia 2014-2020 Program), by grant ED431G 2019/01. RFC is supported by the grant TIN2017-89244-R from MINECO and the recognition 2017SGR-856 (MACDA) from AGAUR (Generalitat de Catalunya).

References

Alemany-Puig, L. (2019). Edge crossings in linear arrangements: From theory to algorithms and applications (Master thesis). Barcelona School of Informatics.

Aronsson, F. S., Kuhlmann, M., Jelic, V., Östberg, P. (2021). Is cognitive impairment associated with reduced syntactic complexity in writing? Evidence from automated text analysis. *Aphasiology*, *35*(7), 900–913. https://doi.org/10.1080/02687038.2020.1742282

Barthélemy, M. (2018). *Morphogenesis of spatial networks*. Springer. https://doi.org/10.1007/978-3-319-20565-6

Bod, R. (2013). *A new history of the humanities: The search for principles and patterns from antiquity to the present*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199665211.001.0001

Bodirsky, M., Kuhlmann, M., Möhl, M. (2005). Well-nested drawings as models of syntactic structure. *10th Conference on Formal Grammar and 9th Meeting on Mathematics of Language*, 195–203.

Bresnan, J. (2000). *Lexical-functional syntax*. Blackwell.

Cayley, A. (1889). A theorem on trees. *Quart. J. Math*, *23*, 376–378. https://doi.org/10.1017/CBO9780511703799.010

Chen, D., Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740–750. https://doi.org/10.3115/v1/D14-1082

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Christiansen, M. H., Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*, 157–205. https://doi.org/10.1016/S0364-0213(99)00003-8

Christiansen, M. H., Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, *31*, 489–558. https://doi.org/10.1017/S0140525X08004998

Corro, C., Le Roux, J., Lacroix, M., Rozenknop, A., Wolfler Calvo, R. (2016). Dependency parsing with bounded block degree and well-nestedness via Lagrangian relaxation and branch-and-bound. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 355–366. https://doi.org/10.18653/v1/P16-1034

Covington, M. A. (2001). A fundamental algorithm for dependency parsing. *Proceedings of the 39th Annual ACM Southeast Conference*, 95–102.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., Manning, C. D. (2014). Universal Stanford dependencies: A cross-linguistic typology. In N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 4585–4592). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1062_Paper.pdf

de Marneffe, M.-C., Manning, C. D. (2008). The Stanford typed dependencies representation. *COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, 1–8. http://aclweb.org/anthology/W08-1301

Dyer, C., Ballesteros, M., Ling, W., Matthews, A., Smith, N. A. (2015). Transition-based dependency parsing with stack long short-term memory. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 334–343. https://doi.org/10.3115/v1/P15-1033

Dyer, C., Kuncoro, A., Ballesteros, M., Smith, N. A. (2016). Recurrent neural network grammars. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 199–209. https://doi.org/10.18653/v1/N16-1024

Eisner, J., Smith, N. A. (2010). Favor short dependencies: Parsing with soft and hard constraints on dependency length. In H. Bunt, P. Merlo, J. Nivre (Eds.), *Trends in parsing technology: Dependency parsing, domain adaptation, and deep parsing* (pp. 121–150). Springer. http://cs.jhu.edu/~jason/papers/#eisner-smith-2010-iwptbook

Esteban, J. L., Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016). The scaling of the minimum sum of edge lengths in uniformly random trees. *Journal of Statistical Mechanics*, 063401. https://doi.org/10.1088/1742-5468/2016/06/063401

Ferrer-i-Cancho, R., C. Gómez-Rodríguez, J. L. E., Alemany-Puig, L. (2022). Optimality of syntactic dependency distances. *Physical Review E*, *105*(1), 014308. https://doi.org/10.1103/PhysRevE.105.014308

Ferrer-i-Cancho, R. (2004). Euclidean distance between syntactically linked words. *Physical Review E*, *70*, 056135. https://doi.org/10.1103/PhysRevE.70.056135

Ferrer-i-Cancho, R. (2006). Why do syntactic links not cross? *Europhysics Letters*, *76*(6), 1228–1235. https://doi.org/10.1209/epl/i2006-10406-0

Ferrer-i-Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, *11*(3), 393–414. https://doi.org/10.1142/S0219525908001702

Ferrer-i-Cancho, R. (2013). Hubiness, length, crossings and their relationships in dependency trees. *Glottometrics*, *25*, 1–21.

Ferrer-i-Cancho, R. (2014). A stronger null hypothesis for crossing dependencies. *Europhysics Letters*, *108*(5), 58003. https://doi.org/10.1209/0295-5075/108/58003

Ferrer-i-Cancho, R., Gómez-Rodríguez, C., Esteban, J. L. (2018). Are crossing dependencies really scarce? *Physica A: Statistical Mechanics and its Applications*, *493*, 311–329. https://doi.org/10.1016/j.physa.2017.10.048

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016a). Crossings as a side effect of dependency lengths. *Complexity*, *21*, 320–328. https://doi.org/10.1002/cplx.21810

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2016b). Liberating language research from dogmas of the 20th century. *Glottometrics*, *33*, 33–34.

Ferrer-i-Cancho, R., Gómez-Rodríguez, C. (2021). Anti dependency distance minimization in short sequences. A graph theoretic approach. *Journal of Quantitative Linguistics*, *28*(1), 50–76. https://doi.org/10.1080/09296174.2019.1645547

Ferrer-i-Cancho, R., Liu, H. (2014). The risks of mixing dependency lengths from sequences of different length. *Glottotheory*, *5*, 143–155. https://doi.org/10.1515/glot-2014-0014

Futrell, R., Levy, R. P., Gibson, E. (2020). Dependency locality as an explanatory principle for word order. *Language*, *96*(2), 371–412. https://doi.org/10.1353/lan.2020.0024

Futrell, R., Mahowald, K., Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, *112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gaifman, H. (1965). Dependency systems and phrase-structure systems. *Information and Control*, *8*, 304–337. https://doi.org/10.1016/S0019-9958(65)90232-9

Gildea, D., Temperley, D. (2010). Do grammars minimize dependency length? *Cognitive Science*, *34*(2), 286–310. https://doi.org/10.1111/j.1551-6709.2009.01073.x

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*, 447–474. https://doi.org/10.1016/S0019-9958(67)91165-5

Gómez-Rodríguez, C. (2016). Restricted non-projectivity: Coverage vs. efficiency. *Computational Linguistics*, *42*(4), 809–817. https://doi.org/10.1162/COLI_a_00267

Gómez-Rodríguez, C., Carroll, J., Weir, D. (2011). Dependency parsing schemata and mildly non-projective dependency parsing. *Computational Linguistics*, *37*(3), 541–586. https://doi.org/10.1162/COLI_a_00060

Gómez-Rodríguez, C., Ferrer-i-Cancho, R. (2017). Scarcity of crossing dependencies: A direct outcome of a specific constraint? *Physical Review E*, *96*, 062304. https://doi.org/10.1103/PhysRevE.96.062304

Gómez-Rodríguez, C., Nivre, J. (2010). A transition-based parser for 2-planar dependency structures. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1492–1501. http://portal.acm.org/citation.cfm?id=1858681.1858832

Gómez-Rodríguez, C., Shi, T., Lee, L. (2018). Global transition-based non-projective dependency parsing. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2663–2674. https://doi.org/10.18653/v1/P18-1248

Gómez-Rodríguez, C., Weir, D., Carroll, J. (2009). Parsing mildly non-projective dependency structures. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*, 291–299. https://doi.org/10.3115/1609067.1609099

Groß, T., Osborne, T. (2009). Toward a practical dependency grammar theory of discontinuities. *SKY Journal of Linguistics*, *22*, 43–90.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková-Razímová, M., Urešová, Z. (2006). Prague dependency treebank 2.0.

Hajicova, E. (1995). Prague school syntax and semantics. In E. Koerner R. Asher (Eds.), *Concise history of the language sciences* (pp. 253–262). Pergamon. https://doi.org/10.1016/B978-0-08-042580-1.50045-3

Hauser, M. D., Chomsky, N., Fitch, W. T. (2002). The faculty of language: What is it, who has it and how did it evolve? *Science*, *298*, 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199252695.001.0001

Hogg, R. V., Tanis, E. A. (1997). *Probability and statistical inference* (7th). Prentice Hall.

Jing, Y., Blasi, D. E., Bickel, B. (2021). Dependency length minimization and its limits: A possible role for a probabilistic version of the Final-Over-Final Condition. *Language*, in press.

Kahane, S., Mazziotta, N. (2015). Syntactic polygraphs. a formalism extending both constituency and dependency. *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, 152–164. https://doi.org/10.3115/v1/W15-2313

Kesavan, H. K. (2009). Jaynes' maximum entropy principle. In C. A. Floudas P. M. Pardalos (Eds.), *Encyclopedia of optimization* (pp. 1779–1782). Springer US. https://doi.org/10.1007/978-0-387-74759-0_312

Klein, D., Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 423–430. https://doi.org/10.3115/1075096.1075150

Köhler, R., Kelih, E., Goebl, H. (2021). Gabriel Altmann (1931–2020). *Journal of Quantitative Linguistics*, *28*(2), 187–193. https://doi.org/10.1080/09296174.2021.1902057

Kübler, S., McDonald, R., Nivre, J. (2009). Dependency parsing. *Synthesis Lectures on Human Language Technologies*, *1*(1), 1–127.

Kuhlmann, M. (2010). *Dependency structures and lexicalized grammars. an algebraic approach* (Vol. 6270). Springer. https://doi.org/10.1007/978-3-642-14568-1

Kuhlmann, M., Jonsson, P. (2015). Parsing to noncrossing dependency graphs. *Transactions of the Association for Computational Linguistics*, *3*, 559–570. https://doi.org/10.1162/tacl_a_00158

Kuhlmann, M., Nivre, J. (2006). Mildly non-projective dependency structures. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, 507–514.

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

Liu, H., Xu, C., Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. *Physics of Life Reviews*, *21*, 171–193. https://doi.org/10.1016/j.plrev.2017.03.002

Miller, P. (2000). *Strong generative capacity: The semantics of linguistic formalism*. Cambridge University Press.

Morrill, G. (2010). *Categorial grammar: Logical syntax, semantics, and processing*. Oxford University Press.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, 149–160.

Nivre, J. (2004). Incrementality in deterministic dependency parsing. *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together (ACL)*, 50–57. https://aclanthology.org/W04-0308/

Nivre, J. (2005). *Dependency grammar and dependency parsing* (tech. rep. MSI 05133). Växjö University, School of Mathematics and Systems Engineering. http://stp.lingfil.uu.se/~nivre/docs/05133.pdf

Nivre, J. (2009). Non-projective dependency parsing in expected linear time. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, 351–359. https://aclanthology.org/P09-1040

Nivre, J., Abrams, M., Agić, Ž., et al. (2019). Universal dependencies 2.4 [LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-2988

Osborne, T., Putnam, M., Gross, T. (2011). Bare phrase structure, label-less trees, and specifier-less syntax: Is minimalism becoming a dependency grammar? *The Linguistic Review*, *28*, 315–364. https://doi.org/10.1515/tlir.2011.009

Ouyang, J., Jiang, J. (2018). Can the probability distribution of dependency distance measure language proficiency of second language learners? *Journal of Quantitative Linguistics*, *25*(4), 295–313. https://doi.org/10.1080/09296174.2017.1373991

Park, Y. A., Levy, R. (2009). Minimal-length linearizations for mildly context-sensitive dependency trees. *Proceedings of the 10th Annual Meeting of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) conference*, 335–343. https://aclanthology.org/N09-1038

Passarotti, M. C. (2016). How far is Stanford from Prague (and vice versa)? Comparing two dependency-based annotation schemes by network analysis. *L'analisi Linguistica e Letteraria*, *1*, 21–46.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. In M. H. Christiansen S. Kirby (Eds.), *Language evolution* (pp. 16–37). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199244843.003.0002

Pitler, E. (2014). A crossing-sensitive third-order factorization for dependency parsing. *Transactions of the Association for Computational Linguistics*, *2*, 41–54. https://doi.org/10.1162/tacl_a_00164

Pitler, E., Kannan, S., Marcus, M. (2013). Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics*, *1*, 13–24. https://doi.org/10.1162/tacl_a_00206

Prüfer, H. (1918). Neuer Beweis eines Satzes über Permutationen. *Arch. Math. Phys*, *27*, 742–744.

Roark, B., Mitchell, M., Hosom, J., Hollingshead, K., Kaye, J. (2011). Spoken language derived measures for detecting mild cognitive impairment. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2081–2090. https://doi.org/10.1109/TASL.2011.2112351

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In N. C. (Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)* (pp. 2334–2341). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/915_Paper.pdf

Sgall, P. (1969). *A functional approach to syntax in generative description of language*. Elsevier.

Smith, N. A., Eisner, J. (2006). Annealing structural bias in multilingual weighted grammar induction. *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, 569–576. https://doi.org/10.3115/1220175.1220247

Temperley, D., Gildea, D. (2018). Minimizing syntactic dependency lengths: Typological/Cognitive universal? *Annual Review of Linguistics*, *4*(1), 67–80. https://doi.org/10.1146/annurev-linguistics-011817-045617

Tomasello, M. (2005). *Constructing a language. A usage-based theory of language acquisition*. Harvard University Press.

Yadav, H., Husain, S., Futrell, R. (2019). Are formal restrictions on crossing dependencies epiphenominal? *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, 2–12. https://doi.org/10.18653/v1/W19-7802

Yadav, H., Husain, S., Futrell, R. (2021). Do dependency lengths explain constraints on crossing dependencies? *Linguistics Vanguard*, *7*(s3), 20190070. https://doi.org/10.1515/lingvan-2019-0070

Yadav, H., Husain, S., Futrell, R. (2022). Assessing corpus evidence for formal and psycholinguistic constraints on nonprojectivity. *Computational Linguistics*, 1–27. https://doi.org/10.1162/coli_a_00437

Yli-Jyrä, A., Gómez-Rodríguez, C. (2017). Generic axiomatization of families of noncrossing graphs in dependency parsing. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1745–1755. https://doi.org/10.18653/v1/P17-1160

Yuan, J., Lin, Q., Lee, J. S. Y. (2021). Discourse tree structure and dependency distance in EFL writing. *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, 105–115. https://aclanthology.org/2021.tlt-1.10