Some word order biases from limited brain resources. A mathematical approach.
Running tittle: Word order biases.

Ramon Ferrer-i-Cancho

Departament de Física Fonamental. Universitat de Barcelona. Martí i Franquès 1, 08028 Barcelona, Spain.
ramon.ferrericancho@gmail.com
Current address:
Departament de Llenguatges i Sistemes Informàtics, TALP Research Center, Universitat Politècnica de Catalunya, Campus Nord, Edifici Ω Jordi Girona Salgado 1-3. 08034 Barcelona, Spain.

Here we propose a mathematical framework for studying word order optimization. The framework relies on the well-known positive correlation between cognitive cost and the Euclidean distance between the elements (e.g. words) involved in a syntactic link. We study the conditions under which a certain word order is more economical than an alternative word order by proposing a mathematical approach. We apply our methodology to two different cases: (a) the ordering of subject (S), verb (V) and object (O) and (b) the covering of a root word by a syntactic link. As for the former, we find that SVO and its symmetric, OVS, are more economical than OVS, SOV, VOS and VSO at least 2/3 of the times. As for the latter, we find that uncovering the root word is more economical than covering it at least 1/2 of the times. With the help of our framework, one can explain some Greenbergian universals. Our findings provide further theoretical support for the hypothesis that the limited resources of the brain introduce biases towards certain word orders. Our theoretical findings could inspire or illuminate future psycholinguistics or corpus linguistics studies.

Keywords: word order, cognitive cost, syntactic dependency, linguistic universals, human language.

1. Introduction

Word order has been the subject of a large amount of research within various fields: typology of language universals [15, 5], psycholinguistics [24, 12] and the evolution of language [22]. A paradigmatic example of word order problem is the sequential arrangement of subject (S), verb (V) and object (O). Hereafter, S, V and O stand for subject, verb and object, respectively. Work on the intersection among generative syntax, typology of linguistics universals and psycholinguistics has placed considerations about the limited resources of our brain [24], at the core syntactic theory [20]. It has been argued that at least some of the basic grammar

conventions of languages ease the processing of sentences [20] or that the ease with which a sentence is processed depends on the distance between syntactically related elements [12, 20]

The focus of this article is surface word order [15, 5], namely, the word order that real sentences show. In particular, we will study the word orders that are in overall more economical, i.e. consume less brain resources. The emphasis of this article is on using a simple but powerful mathematical approach.

Word order has been studied from a mathematical and a computational perspective. The literature is large and disperse so we just review some representative works from different perspectives. Quantitative linguistics studies suggest that the most frequent items tend to appear first in the sentence [7]. Computational experiments suggest that word orders may naturally emerge in a population of interacting agents without the need of selection of the fittest orders [22, 23]. The ordering of the triple (S, V, O) that emerges differs from run to run. Besides, the contribution of word order to grammar efficiency [30] or the number of unambiguous word orders that can be generated [29] has also been studied mathematically. In none of these works, the cost derived from the distance between syntactically related items is considered.

The remainder of the article is organized as follows. Section 2 introduces the syntactic formalism that we will use for studying word order (i.e. dependency grammar). Section 3 presents the factor that determines word order in our word order idealization (i.e. the Euclidean distance minimization between syntactically related words). Section 4 presents the kind of mathematical approach we will use for determining the best word order according to the factor mentioned above. Two applications of our mathematical approach to the bias for SVO order and the tendency to not cover the root of a sentence with a syntactic link are provided in Section 5. Section 6 shows how the mathematical framework could be applied to research on linguistic universals. The article ends with a discussion in Section 7.

2. Dependency grammar

Dependency grammar is a class of grammatical formalisms [26, 21, 28] specifying how pairs of words link in sentences. Typically, two words are linked if one syntactically depends on the other. Links are syntactic dependencies. Most links are directed and the arc goes from the head word to its modifier or vice versa depending on the convention used. Head and modifier are primitive concepts in the dependency grammar formalism (Fig. 1). The dependency grammar formalism distinguishes some cases, such as coordination, where there is no clear direction [27]. In the examples used here arcs go from the head to its modifier, but link direction is not relevant here because we are only concerned about the distance between linked words.
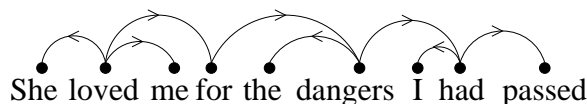
Fig. 1. The syntactic structure of the sentence 'She loved me for the dangers I had passed' following the conventions in [26]. Vertices are words and the arcs stand for syntactic dependencies. Following the conventions in [26], arcs go from a head to its modifier. The pronoun 'she' and the verb 'loved' are syntactically dependent in the sentence. 'she' is the modifier of the verbal form 'loved', which is its head. Similarly, the action of 'loved' is modified by its object 'me'. 'loved' is the root vertex.

3. Euclidean distance minimization

The distance between syntactically related items in sentences is a basic ingredient of the cost of a sentence [18, 13, 20] and has been used for explaining surface word order tendencies [20]. Here we focus on the Euclidean distance between syntactically linked words in sentences. Now we are going to explain our definition of distance precisely. Here we assume that words are placed on a straight line following the order of a sentence (as in Fig. 1). Our convention consists of assigning position one to the first word of the sentence and adding one after every word for calculating the positions of the next word (hence 'she' has position 1, 'loved' has position 2 and so on). We define $\pi(v)$ as the position of word v and the Euclidean distance between two words, u and v, is defined as $d(u,v) = |\pi(u) - \pi(v)|$, so $d(u,v) = d(v,u)$. The units of distance measure are words (notice that we could have calculated the distance in syllables, for instance). We are only interested in the distance between directly connected words. Table 1 lists the positions of every word and the distance to the sender of the arc for the sentence in Fig. 1 (the dependency grammar formalism generally assumes that every vertex receives one arc except for the root word, that receives no arc). If the word 'she' was moved to the end of the sentence, then all distances in Table 1 would remain the same except for d('she', 'loved') = 8 (words). Hereafter we assume that distance in measured in words by default.

It has been shown that the distance between syntactically related words is significantly small and that constraining the distance between syntactically related words while maximizing the occupation of all possible distances can explain the exponential trend of that distance [8]. Cost minimization, or equivalently least effort, is a key principle for explaining universals in quantitative linguistics. For instance, Zipf's law [31] for word frequencies can be explained by minimizing the cost of word use and maximizing the information conveyed by words [11, 9]. Another example is the fact that the rarity of syntactic dependency crossings can be explained by minimizing the cost of syntactic dependency links [10]. Here we assume that distance minimization is a key factor for understanding word order tendencies as in [20].

4   *Ramon Ferrer-i-Cancho*

word	$\pi(word)$	sender	$d(word, sender)$
she	1	*loved*	1
loved	2	—	—
me	3	*loved*	1
for	4	*loved*	2
the	5	*dangers*	2
dangers	6	*for*	2
I	7	*had*	1
had	8	*dangers*	2
passed	9	*had*	1

Table 1. Every word or the sentence 'She loved me for the dangers I had passed', the position of every word ($\pi(word)$) and the distance (in words) of every word to the sender of arc ($d(word, sender)$).

4. Minimum linear arrangement

Minimizing the sum of the Euclidean distances between linked vertices on a network where vertices follow a sequence is known as the minimum linear arrangement (m.l.a.) problem in computer science [4]. More formally, suppose that we have a network whose set of vertices is $\mathcal{V}$ and its set of arcs is $\mathcal{A}$ (a directed graph). In our case, $\mathcal{V}$ contains the words of a sentence and $\mathcal{A}$ the syntactic dependency links of this sentence. π defines a minimum linear arrangement if

$$\Omega(\pi, \mathcal{A}) = \sum_{(u,v) \in \mathcal{A}} d(u,v) \tag{1}$$

is minimum.

The fact that human utterances are linear (i.e. a sequence of basic units) was early emphasized by the French linguist Ferdinand de Saussure [3]. Speaking implies transforming commonly multidimensional thoughts into the single dimension of speech. The memory expenditure of this process depends on the word order chosen [24]. Linearization implies solving an optimization problem.

Before we proceed we need to warn the reader. The mathematical framework we will introduce here is an idealization of the real surface word order problem. Our strategy is similar to that followed in physics for gases. One imagines an ideal gas (where collisions between molecules can be neglected to some extent) and then one studies how much can be predicted by this idealization. The idealized gas turns out to have a satisfactory predictive power in many applications. The point is starting from a simple model and then moving towards a more complicated one, not in the opposite way. This is important because our model does not consider all the factors possible involved in the word order problem. Eventually, the success of our selected idealization can only be supported by the successful predictions it can make, as in the case of gases in physics. First, our framework assumes an ideal speaker whose surface word order selections are not influenced by the order selections made by

other individuals in his environment. Second, our framework assumes that the only force involved in selecting a certain surface word order is minimizing the distance between syntactically related words (this distance is explained in full detail later on), which is equivalent to saying that there maybe other factors, e.g., the tendency of topical information to generally occur early in a sentence [25, 14], but they can be neglected. This idealization is congenial with the hypothesis that free word order is not driven by pragmatic considerations, as it is generally believed, but by the need to recognize syntactic structures rapidly online [20]. More formally, we assume that minimizing $\Omega(\pi, \mathcal{A})$ is the major force in determining π, the ordering of words of a sentence. Indeed, we assume a particular case of m.l.a. where only constituents at a certain level can be rearranged while keeping their internal ordering remains the same. Third, our framework assumes that syntactic links do not cross. Indeed, this is not a heavy assumption since this property is a natural consequence of the minimization of the length of syntactic links [10]. The only simplification we make with this regard is assuming that crossings never appear whereas they are found rather exceptionally in actual sentences. Fourth, our framework assumes that measuring Euclidean distance in words is precise enough at least for the purpose of this article. One could measure this distance in syllables or phonemes (it would still be a discrete number) or even in elapsed time (then it would become a continuous amount). Finally, it is important to call the attention of the reader on the fact that idealization is not unique to this article among studies of word order universals. In recent studies of the ordering of S, V and O [5], the dominant word order of a language is not defined as the most frequent word order globally but the most frequent word order of declarative sentences (not interrogatives or exclamatives) with the further constraint that S and the O cannot be a pronoun (our approach to the ordering of S, V and O that we will introduce in this article is not limited to these particular cases).

5. Case studies

We introduce some mathematical notation that will be used in the next subsections. A constituent x is a set of vertices, $x \subseteq \mathcal{V}$. For instance, $S_1 = \{\text{'she'}\}$ and $S_2 = \{\text{'the', 'dangers'}\}$ are, respectively, the subjects of the main clause and the secondary clause of the sentence in Fig. 1. We define A_x as the set of all the edges formed between vertices of the set x ($x \subseteq V$ and $A_x \subseteq A$). Technically, A_x contains the edges of the subtree induced by the vertices in x [1]. We define Ω_x as the sum of the Euclidean distances of the edges of A_x. Technically, Ω_x is the abbreviation for $\Omega(\pi, A_x)$. We define r_x as the root of the tree (or subtree) formed by the vertices of x (r_x may not be defined if the vertices of x do not form a tree). For instance, $r_{S_1} = \text{'she'}$ and $r_{S_2} = \text{'dangers'}$. Linguistically, r_x is the head of constituent x. We assume that constituents cannot be empty (there is at least a head). The length in words of a constituent x can be written as $|x| = L_x + 1 + R_x$, where $|...|$ is the cardinality operator and L_x and R_x are, respectively, the number of vertices to the
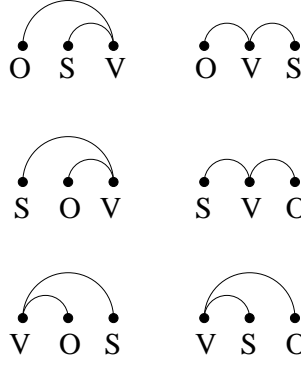
6 *Ramon Ferrer-i-Cancho*

Fig. 2. All the possible orderings of subject (S), verb (V) and and object (O). We assume that the length of the sentence (in words) is $n = 3$ (and there are no empty constituents) or equivalently, that S, V and O contain a single word, represented by a filled black circle as in Fig. 1. The cost of each ordering is $\Omega = 3$, except for SVO and OVS, whose cost is $\Omega = 2$.

left and to the right of r_x within x. More formally,

$$R_x = |\{v|\pi(v) > \pi(r_x) \text{ and } v \in x|\} \qquad (2)$$

$$L_x = |\{v|\pi(v) < \pi(r_x) \text{ and } v \in x|\}. \qquad (3)$$

Therefore, $L_{S_1} = R_{S_1} = 0$ and $|S_1| = 1$ while $L_{S_2} = 1$, $R_{S_2} = 0$ and $|S_2| = 2$. We define n as the length in words of the sentence ($n = 9$ in Fig. 1). We have $n = |\mathcal{V}|$.

5.1. *Case study I: ordering of S, V, and O.*

.

Here we focus on the particular problem of minimizing Ω for the triple (S, V, O). Here S, V and O stand for the set of words (or vertices) in the subject, verb and object constituents. An m.l.a. can very easily explain why SVO or its symmetric, i.e. OVS, should be preferred as surface word order when S, V and O are formed by just one word each, namely, $|S| = |V| = |O| = 1$ (see Fig. 2 for all the possible orderings satisfying this condition). In this case, $A = \{(r_V, r_S), (r_S, r_O)\}$. We extend the definition of Ω to allow it to take w, a permutation of S, V and O instead of the argument π. If the permutation is the sequence x, y, z (where $x \in \{S, V, O\}$ and $x \neq y \neq z$), $\pi(r_x) = 1$, $\pi(r_y) = 2$ and $\pi(r_z) = 3$. Given a certain permutation one can derive a unique π. Thus, we define $\Omega'(w, \mathcal{A})$ as the value of $\Omega(\pi, \mathcal{A})$ for the π that derives from w. $|S| = |V| = |O| = 1$ gives

$$\Omega'(w, \mathcal{A}) = \begin{cases} 2 \text{ if } w \in \{SVO, OVS\} \\ 3 \text{ otherwise.} \end{cases} \qquad (4)$$

The proportion of orders of S, V, O where the order SVO or OVS gives the smallest Ω (including SVO themselves) is $2/3$ and the proportion of orders where the orders SVO or OVS give the smallest Ω including SVO and OVS themselves) is one.
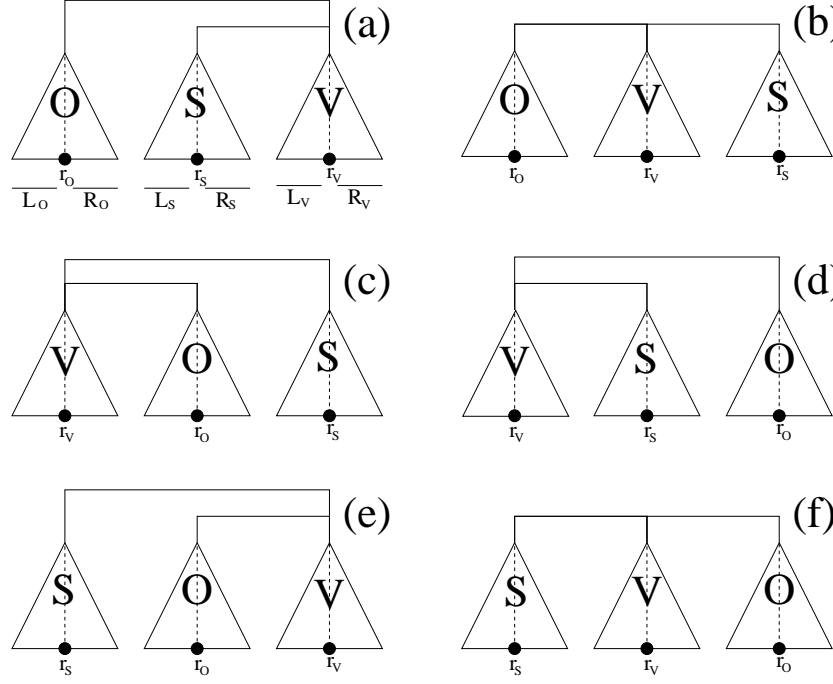
Fig. 3. Scheme of a sentence made of object (O), subject (S) and verb (V). r_x is the root of the subtree x, where $x \in \{O, S, V\}$. L_x and R_x are, respectively, the number of words of x to the left and the right of r_x within x. Again, $x \in \{O, S, V\}$. If the sentence has n words then $L_S + R_S + L_V + R_V + L_O + R_O + 3 = n$.

We want to consider the general case where S, V and O can be made of more than one word, i.e. $|S|, |V|, |O| \geq 1$. The whole sentence is formed by linking r_V with r_S and r_O as in Fig. 3. Assuming that Ω_S, Ω_V and Ω_O do not depend on the type of (S, V, O) arrangement, we may write

$$\delta_x = \Omega_x - \Omega_S - \Omega_V - \Omega_O, \tag{5}$$

where x is   whatever (S, V, O) arrangement,   i.e. $x \in \{OSV, OVS, SOV, SVO, VOS, VSO\}$. It can be easily seen from Eq. 3 that SVO is more economical than a word order x if $\Omega_{SVO} < \Omega_x$.

The condition $\Omega_{SVO} < \Omega_x$ becomes $\delta_{SVO} < \delta_x$ for $x \neq SVO$. We have

$$\begin{aligned}
\delta_{SVO} &= R_S + L_V + R_V + L_O + 2 \\
\delta_{SOV} &= 2L_V + 2R_O + L_O + R_S + 3 \\
\delta_{VSO} &= 2R_V + 2L_S + L_O + R_S + 3 \\
\delta_{OSV} &= 2R_S + 2L_V + R_O + L_S + 3 \\
\delta_{VOS} &= 2R_V + 2L_O + R_0 + L_S + 3 \\
\delta_{OVS} &= R_O + L_V + R_V + L_S + 2.
\end{aligned} \tag{6}$$

8 *Ramon Ferrer-i-Cancho*

We define $\Delta_x = R_x - L_x$ with $x \in \{S, V, O\}$. The condition $\delta_{SVO} < \delta_{SOV}$ leads to

$$\Delta_V < 2R_O + 1. \tag{7}$$

The condition $\delta_{SVO} < \delta_{VSO}$ leads to

$$-\Delta_V < 2L_S + 1. \tag{8}$$

The condition $\delta_{SVO} < \delta_{OSV}$ leads to

$$\Delta_V < \Delta_O + L_S + 1. \tag{9}$$

The condition $\delta_{SVO} < \delta_{VOS}$ leads to

$$\Delta_S < \Delta_V + R_O + 1. \tag{10}$$

Finally, the condition $\delta_{SVO} < \delta_{OVS}$ leads to

$$\Delta_S < \Delta_O. \tag{11}$$

Notice that we are not studying a full m.l.a. of S, V and O because we treat S, V and O as undecomposable blocks. We do not allow one to reorder the words within a certain block, e.g., swapping the words before r_S (the head of S) and the word after r_S. A full m.l.a. can only take place in our framework when $n = 3$ (recall that n is the length of the sentence in words)

Eqs. 7, 8, 9, 10 and 11 suggest the existence of three classes of (S, V, O) orders relative to SVO order:

Class I: SOV and VSO
This is the class of orders where O follows S immediately. Notice that Eq. 7 and 8 follow the same template. These equations differ only in the sign of Δ_V and the remaining parameter (R_O or L_S) involved in the right hand side (r.h.s). Class I word orders have some interesting properties. First, in both Class I word orders, Δ_V is a key quantity. SVO is more economical than SOV (i.e. Eq. 7 is satisfied) when $\Delta_V \leq 0$ since $R_O \geq 0$. In contrast, SVO is more economical than VSO (i.e. Eq. 8 is satisfied) when $\Delta_V \geq 0$ since $L_S \geq 0$. Moreover, the lower cost of SVO over SOV implies, in some circumstances, the lower cost of SVO over VSO. Vice-versa, the lower cost of SVO over VSO implies, in some circumstances, the lower cost of SVO over SOV. To see it in detail, if Eq. 7 is satisfied with $\Delta_V \geq 0$, then Eq. 8 is trivially satisfied. Inversely, if Eq. 8 is satisfied with $\Delta_V \leq 0$, then Eq. 7 is trivially satisfied. A particular case is $\Delta_V = 0$, which implies that SVO is more economical than any Class I word order. For this reason, if V is made of a single word (i.e. the sentence has a single word verb and the remainder of the words fall in either S or O) then SVO is more economical no mater how the other constituents are made.

Class II: OSV and VOS
This is the class of word orders where S follows O immediately. Notice that Eqs. 9 and 10 follow the same template. These equations differ only in the sign of Δ_V, the

Order	Frequency
SOV	497
SVO	435
VSO	85
VOS	26
OVS	9
OSV	4
No dominant order	172
Total	1228

Table 2. Frequency of a word orders in world languages. Borrowed from [5].

constituent x involved in the Δ_x of the l.h.s. and the remaining parameter (R_O or L_S again) involved in the r.h.s. Within this class, Δ_V is a key quantity but other parameters are need to get configurations where SVO is more economical with regard to Class I. The condition $\Delta_V = \Delta_S$ gives that SVO is more economical than VOS (recall Eq. 10) whereas the condition $\Delta_V = \Delta_O$ gives that SVO is more economical than OSV (recall Eq. 9). In particular, if V and S are made of a single word then the lower cost of SVO over VOS follows trivially regardless of how O is made of. Similarly, if V and O are made of a single word ($\Delta_V = \Delta_S = 0$) then the lower cost of SVO over OSV follows trivially regardless of the composition of S.

Class III: OVS
This is the class of the reverse of SVO.

We investigate the relationship between the six possible orderings of S, V and O and the three classes. Fig. 4 shows the network of permutations of S, V and O. Two permutations are connected if one gives the other by swapping two consecutive constituents or vice-versa. The network is a ring made of six vertices. Interestingly, the two most frequent dominant word orders (Table 2), SVO and SOV are consecutive in the permutation network. We aim to study if the consecutive placement of this pair of word orders can be explained by chance (the null hypothesis) or by a special factor relying on the permutation ring (the alternative hypothesis). We define k as the number of word orders of a ring. Assuming that the k orders are equally likely, the probability that two randomly chosen orders are adjacent in the permutation ring is

$$p = \binom{k}{2}^{-1} = \frac{2}{k(k-1)}. \tag{12}$$

In the space of permutations of SVO (where $k = 6$), we obtain $p = 1/15 \approx 0.066$. At a significant level of 0.05, we cannot reject the null hypothesis although p is low.
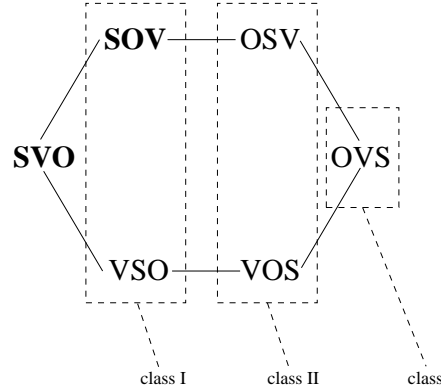
10   *Ramon Ferrer-i-Cancho*

Fig. 4. The permutation of space of S, V and O. A pair of permutations (x,y) is linked if x can give y by swapping two consecutive constituents or vice-versa. Bold face is used for indicating the two most frequent dominant word orders in world languages [5].

L_S	R_S	L_V	R_V	L_0	R_0	Δ_S	Δ_V	Δ_O
0	0	0	0	0	1	0	0	1
0	0	0	0	1	0	0	0	-1
0	0	0	1	0	0	0	1	0
0	0	1	0	0	0	0	-1	0
0	1	0	0	0	0	1	0	0
1	0	0	0	0	0	-1	0	0

Table 3. All the possible configurations that can be formed in sentences of length $n = 4$. $\Delta_x = R_x - L_x$.

We have seen above some special conditions making SVO more economical than any other order. We define n as the length of the sentence in words. Hereafter, length is measured in words. In order to investigate how much better SVO is in general against the remaining orders, we generate all distinct possible configurations of $(L_S, R_S, L_V, R_V, L_O, R_O)$ obeying

$$L_S + R_S + L_V + R_V + L_O + R_O + 3 = n \qquad (13)$$

$$L_S, R_S, L_V, R_V, L_O, R_O \geq 0 \qquad (14)$$

with the help of the computer while keeping n fix. When $n = 3$, there is only one possible configuration, namely $L_S = R_S = L_V = R_V = L_O = R_O = 0$. When $n = 4$, there are only 6 possible configurations characterized by $L_S + R_S + L_V + R_V + L_O + R_O = 1$ (Fig. 3).

We define $p_>^x(n)$ as the proportion of configurations of $(L_S, R_S, L_V, R_V, L_O, R_O)$ where SVO is better than a target word order x, (i.e. the proportion of configurations where SVO has smaller Ω than x, with $x \in \{OSV, OVS, SOV, VOS, VSO\}$)

Class	x	$p_>^x(3)$	$p_\geq^x(3)$	$p_>^x(4)$	$p_\geq^x(4)$	$p_\geq^x(n \to \infty)$, $p_\geq^x(n \to \infty)$
I	SOV	1	1	5/6	1	≈ 0.81
I	VSO	1	1	5/6	1	≈ 0.81
II	OSV	1	1	2/3	1	≈ 0.68
II	VOS	1	1	2/3	1	≈ 0.68
III	OVS	0	1	1/2	5/6	≈ 0.5

Table 4. $p_>^x(n)$ and $p_\geq^x(n)$ for specific values of n. $p_>^x(n)$ is the proportion of configurations where SVO is more economical than the target word order x. $p_\geq^x(n)$ is the proportion of configurations where SVO is more or equally economical than the target word order x. The values of $p_>^x(n)$ and $p_\geq^x(n)$ collapse when $n \to \infty$.

in sentences of length n using Eq. 7,8,9,10 and 11. Similarly, we define $p_\geq^x(n)$ as the proportion of configurations where SVO is better or equal than a target word order x in sentences of length n ($x \in \{OSV, OVS, SOV, VOS, VSO\}$. Table 4 shows $p_>^x(n)$ and $p_\geq^x(n)$ for $n \in \{3, 4\}$ and also approximate values for $n \to \infty$. It is easy to see that $p_>^x(3) = p_\geq^x(3) = 1$ provided $x \notin \{OVS, SVO\}$ when $n = 3$ as in Fig. 2. If $x \in \{OVS, SVO\}$ then $p_>^x(3) = 0$ and $p_\geq^x(3) = 1$.

Fig. 5 shows $p_>^x(n)$ and $p_\geq^x(n)$ for all target word orders x. The curves for members of the same class are identical and differ from one class to another. These results and Table 4 provide support for the a priori arbitrary classification we made based on the algebraic form of the inequalities in Eqs. 7,8,9,10 and 11. In some cases, $p_\geq^x(n)$ has always its maximum and its minimum in the extremes $n = 3$ and $n \to \infty$. In contrast, $p_>^x(n)$ has minima far from the extremes for Class I and II. We have that $p_>^x(n) \geq 2/3$ if x is of Class I or II. In this case, the lower bound of $p_>^x(n)$ comes from $p_>^x(4) = 2/3$ when x is of Class II (recall Table 4). When x is of Class I, $p_>^x(n)$ is minimized for $n \in \{7, 8\}$. In this class, $p_>^x(n) \geq p_>^x(7) = p_>^x(8) \approx 0.801$.

In sum, all different configuration of $(L_S, R_S, L_V, R_V, L_O, R_O)$ are equally likely (roughly speaking, all these configurations have the same frequency or have the same weight) then the complexity of the lower cost of SVO reduces from five alternative word orders to just three classes.

So far have considered only the advantage of SVO over the remaining word orders in sentences of arbitrary length. Being OVS the symmetric of SVO, positive biases towards OVS (measured as the proportion of configurations where OVS is more economical that a target word order) can also be obtained with the same procedure used for SVO. Being OVS the order of only 0.95% of world languages with a dominant word order (Table 2), we leave the translation of the mathematical framework from SVO to OVS for future work. In the discussion section, we speculate about why OVS is rare despite being as advantageous as SVO.
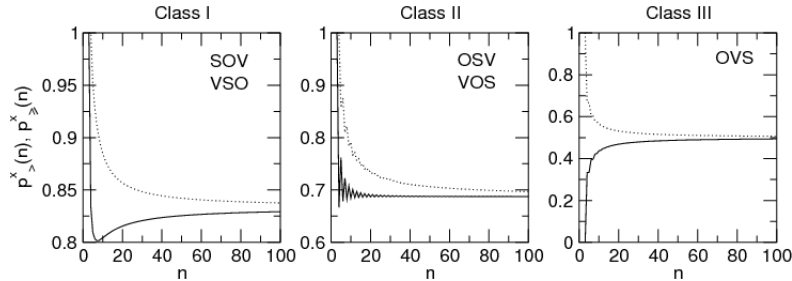
12 *Ramon Ferrer-i-Cancho*

Fig. 5. The economy of SVO versus n, the sentence length. $p_>^x(n)$ (solid line) is the proportion of configurations where SVO is more economical than the target word order x. $p_\geq^x(n)$ (dotted line) is the proportion of configurations where SVO is more or equally economical than the target word order x. Target word orders are grouped into tree classes: Class I for SOV and VSO (left), Class II for OSV and VOS (center) and Class III for OVS (right). The curves for target word orders within the same class are identical.

5.2. *Case study II: government of the top node*

Euclidean link distance minimization can shed light on the origins of projectivity [26], a common property of the syntactic structure of sentences. A sentence is projective if and only if among the arcs of dependency linking its word forms (i) no arc crosses another arc and (ii) no arc governs the top node [26] (the top node of the sentence in Fig. 1 is 'loved'). In this section, we aim to quantify the advantage of syntactic dependency trees that satisfy (ii) over those that do not satisfy it. The structure of the sentence in Fig. 1 obeys (ii) but if we moved "I had passed" to the beginning of the sentence then (ii) would be violated (of course, the sentence would not be proper English; this modification is just made to show how the tree of a sentence that violates (ii) would look like). It has been argued that (i) is a side effect of minimizing the Euclidean distance between linked words [10]. Instead of arguing that (ii) cannot be violated (strong projectivity), we will show that there is a bias towards satisfying (ii) when Euclidean distance minimization works. We will use the same methodology as in Section 5.1.

(ii) is a particular case of vertex covering. A vertex w is covered by the link formed by the pair of vertices (u, v) $(u \neq v)$ if and only if $min(\pi(u), \pi(v)) < \pi(w) < max(\pi(u), \pi(v))$. (ii) is a particular case of a root that is not covered by any edge (our concept of covering neglects arrow directions). Here we will focus on a more general property than (ii). (ii) concerns the top node only), i.e. the covering of a target vertex r_A by an arc.

We assume that the whole tree is made of two subtrees induced by two partitions of the sets of vertices A and B, with $A \cup B = \mathcal{V}$ (Fig. 7 (a)) and whose roots are r_A and r_B, respectively. The trees are connected by an arc from a vertex v in A to r_B. For simplicity, we assume that that the root r_A can be covered by a single edge formed between vertices v and r_B. The edge formed by v and r_B divides the sentence structure into the two subtrees induced by A and B. r_A is the root of the

whole sentence. A is the subtree whose root can be covered by linking the vertex pair (v, r_B).

We define Ω_{COV} and Ω_{UNCOV} as the cost of the whole tree when r_A is covered and when r_A is uncovered, respectively. We could have four different cases depending on whether the root is covered or not and also depending on whether v precedes r_A or follows it (Fig. 7). We define Ω_A and Ω_B as the cost of the subtrees induced by A and B, respectively. We define δ_{COV} and δ_{UNCOV} as the cost of the potentially covering arc in the covered and the uncovered configuration, respectively. We have

$$\Omega_{COV} = \Omega_A + \Omega_B + \delta_{COV} \tag{15}$$

$$\Omega_{UNCOV} = \Omega_A + \Omega_B + \delta_{UNCOV} \tag{16}$$

Condition (ii) (that is, the higher economy of the uncovered configuration) is warranted when $\Omega_{COV} > \Omega_{UNCOV}$, that is when

$$\delta_{COV} > \delta_{UNCOV}. \tag{17}$$

If $\pi(v) < \pi(r_A)$ (Fig. 7 (a)-(b)) then

$$\delta_{COV} = R_A + L_B + \pi(r_A) - \pi(v) + 1 \tag{18}$$

$$\delta_{UNCOV} = L_A + R_B + \pi(v) - \pi(r_A) + 1 \tag{19}$$

and thus Eq. 17 becomes

$$\Delta_A - \Delta_B + 2(\pi(r_A) - \pi(v)) > 0. \tag{20}$$

Recall $\Delta_x = R_x - L_x$, where x is a tree (or subtree).

If $\pi(v) > \pi(r_A)$ (Fig. 7 (c)-(d)) then

$$\delta_{COV} = R_B + L_A + \pi(v) - \pi(r_A) + 1 \tag{21}$$

$$\delta_{UNCOV} = L_B + R_A + \pi(r_A) - \pi(v) + 1 \tag{22}$$

and thus Eq. 17 becomes

$$\Delta_B - \Delta_A + 2(\pi(v) - \pi(r_A)) > 0. \tag{23}$$

If the sentence length is $n = 3$, then we have $\Omega_{COV} = 3$ for covered configurations and $\Omega_{COV} = 2$ for the uncovered configurations (Fig. 6). Thus, uncovered configurations are more economical than covered configurations when $n = 3$.

Now we focus on the case $\pi(v) < \pi(r_A)$ (the following calculations yields identical results for the opposite case). If follows from Eq. 17 that the the cost of a tree is fully specified by the quadruple (L_A, R_A, L_B, R_B), where

$$L_A + R_A + L_B + R_B + 2 = n \tag{24}$$

and

$$L_A \geq \pi(r_A) - \pi(v). \tag{25}$$

We define a well-formed quadruple as a tuple (L_A, R_A, L_B, R_B) such that satisfies Eqs. 24 and 25. We define $q_>^x$ as the proportion of well-formed quadruples
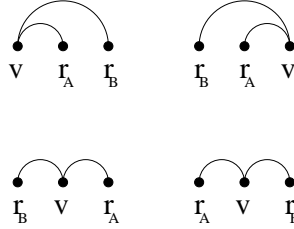
Fig. 6. All the possible configurations of a root r_A and a link (v, r_B) that may cover r_A when the sentence length is $n = 3$.

L_A	R_A	L_B	R_B	Δ_A	Δ_B
1	0	0	1	-1	1
1	0	1	0	-1	-1
1	1	0	0	0	0
2	0	0	0	-2	1

Table 5. All the possible configurations that can be formed in sentences of length $n = 4$. $\Delta_x = R_x - L_x$.

(L_A, R_A, L_B, R_B) where uncovering the root is more economical than covering it with with $x = \pi(r_A) - \pi(v)$. We define $q^x_{\geq}$ as the proportion of well-formed quadruples (L_A, R_A, L_B, R_B) where uncovering the root is more or equally economical than covering the root with $x = \pi(r_A) - \pi(v)$. When $n = 3$, we have that the only well-formed quadruple with $\pi(r_A) - \pi(v) = 1$ is $(1, 0, 0, 0)$, so $q^1_>(3) = q^1_{\geq}(3)$. If $n = 4$, the only well-formed quadruples are shown in Table 5, giving $q^1_>(4) = 1/2$ and $q^1_{\geq}(4) = 1$. Fig. 8 shows $q^x_>(n)$ and $q^X_{\geq}(n)$ for $1 \leq x \leq 4$. We have that $q^x_>(n) \geq 1/2$ (in the domain of n explored in Fig. 8). This means that uncovering the edge is more advantageous at least in $1/2$ of the cases. In Fig. 8, it can also be seen that $q^x_n(n)$ tends to decrease as n grows, suggesting that uncovering configurations should appear more frequently in short sentences than in long sentences, and also that that $q^{x+1}_>(n) > q^x_{\geq}(n)$ and $q^{x+1}_>(n) > q^x_{\geq}(n)$. Thus, separating v from r_A more cannot turn covering configurations more economical as one may intuitively expect.

6. Greenberg's universals revisited

The aim of this section is illustrating how our framework could be employed for research on linguistic universals. We will focus on two Greenbergian universals:

Universal 16. In languages with dominant order VSO, an inflected auxiliary always precedes the main verb. In languages with dominant order SOV, an inflected auxiliary always follows the main verb [16].

Universal 17a With overwhelmingly more than chance frequency, languages with
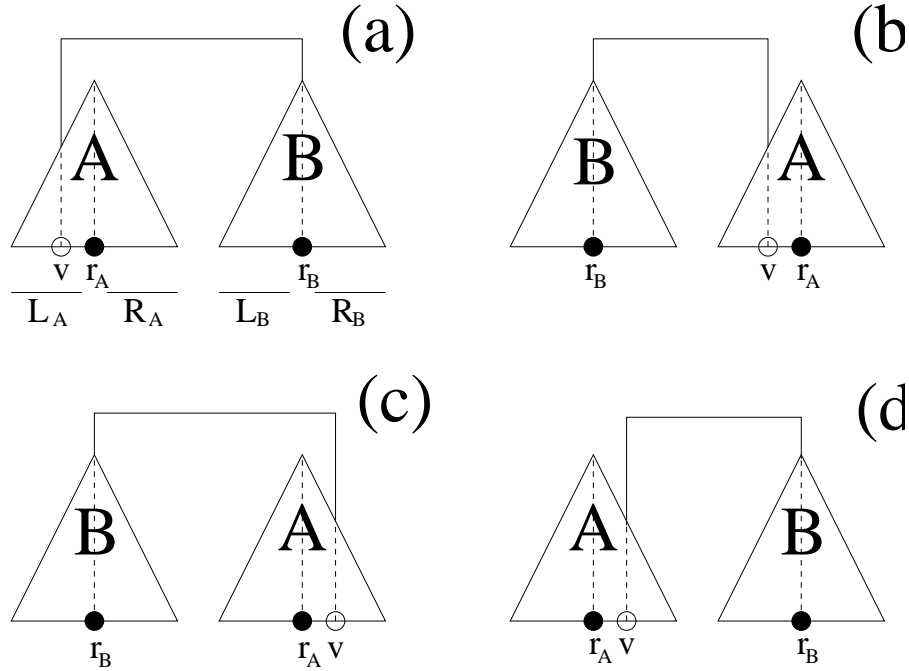
Fig. 7. Scheme of a sentence where, r_A, the root word of the sentence, is covered by an arc linking words v and r_B. The arc divides the sentence structure into two subtrees: A and B. r_A has L_A words on the left and R_A words on the right. Similarly, r_B has L_A words on the left and R_A words on the right. The sentence has n words so $L_A + R_A + L_B + R_B + 2 = n$.

dominant order SVO have the adjective after the noun and languages with dominant order SOV have the adjective before the noun (Appendix). This universal is inspired on Greenberg's universal 17 [16] which is not supported by recent studies (Appendix).

Notice that world Universal 16 and 17a concern word orders of Class I. Let us start with Universal 16. If VSO is dominant then a necessary condition for VSO begin more economical than SVO is the existence of a mechanism for reducing the a priori chance that SVO is a more economical word order, that is, the chance that $-\Delta_V < 2L_S + 1$ (Eq. 8) is satisfied. One way of achieving this is by introducing a bias towards larger values of L_V, which decreases Δ_V. This can be made by putting inflected auxiliaries before the main verb, as stated in Universal 16. Similarly, if SOV is dominant then there must be a mechanism for reducing the a priori chance that SVO is a more economical word order, that is, the chance that $\Delta_V < 2R_O + 1$ (Eq. 7) is satisfied. One way of achieving this is by introducing a bias towards larger values of R_V, which increases Δ_V. This can be made by putting inflected auxiliaries after the main verb, as stated in Universal 16.
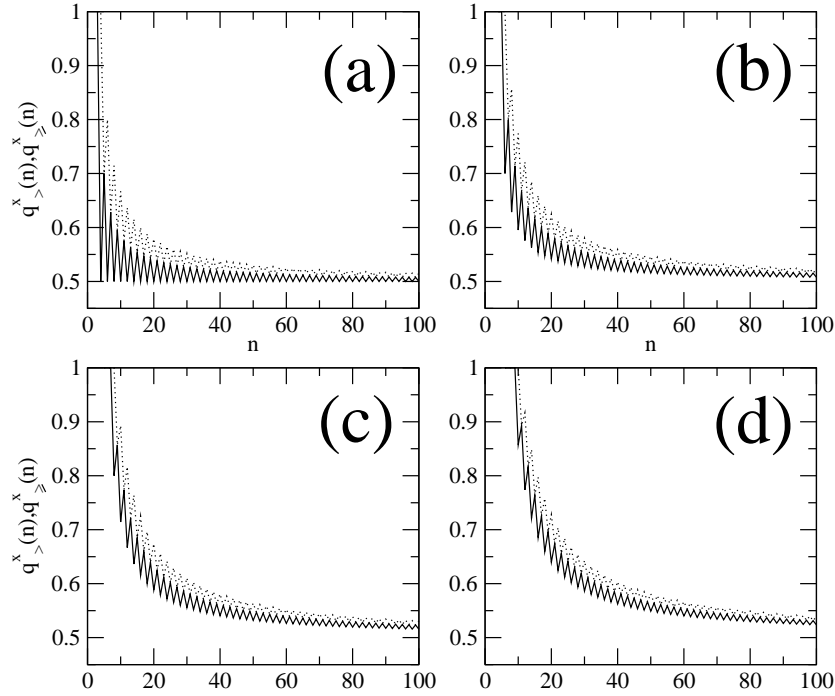
16

Fig. 8. The economy of uncovering the root versus n, the sentence length. $q_>^x(n)$ (solid line) is the proportion of configurations where uncovering the root is more economical than covering it. $q_{\geq}^x(n)$ (dotted line) is the proportion of configurations where uncovering the root is more or equally economical than covering it. x is the distance (in words) between the vertex involved in the edge that may cover the root and lays in the same connected component of the root after removing this edge. (a) $x = 1$. (b) $x = 2$. (c). $x = 3$. (d) $x = 4$.

As for Universal 17a, the dominance of SOV needs that the a priori chance that $\Delta_V < 2R_O + 1$ (Eq. 7) is satisfied is reduced. One way of achieving it is by introducing a bias towards smaller values of R_O. Consistently, languages where SOV is dominant tend to put adjectives before the noun, decreasing R_O. Inversely, keeping SVO more economical than SOV needs that the a priori chance that $\Delta_V < 2R_O + 1$ (Eq. 7) is satisfied is increased. One way of achieving it is by introducing a bias towards larger values of R_O. Consistently, languages where SVO is dominant tend to put adjectives after the noun, increasing R_O.

We need to be conservative and not interpret that Universal 16 and Universal 17a constitute strong support for a distance minimization principle (even if one could show that the involved p-values are 0). The kind of statistical test of correlation used for Universal 16 and Universal 17a cannot determine the actual reason of the correlation [6] . For instance, sharing a certain order could be due convergence from a distance minimization principle or by other kind of factors such as inheritance

from a common evolutionary ancestor or diffusion by contact between neighbouring evolutionary unrelated (or distantly related) languages [6]. Therefore, Universals 16 and 17a are consistent with our mathematical framework based on distance minimization but do not necessary imply strong support for it.

Besides, in order to improve this kind of analysis, one would have to address various questions. For instance, knowing Eqs. 9 and 10, the arguments used for explaining Universal 16 are not only valid for SOV and VSO but also for other word orders placing V at the beginning or at the end: OSV and VOS. Why is there not an equivalent universal for OSV and VOS? Is it just simply due to the fact that OSV and VOS are rare word orders and thus no statistically significant conclusions can be obtained?

7. Discusion

We have proposed a way of measuring the a priori bias towards towards SVO order, or the a priori bias for not covering the root word of a sentence. In particular, we have found that SVO is more economical than OVS, SOV, VOS and VSO at least 2/3 of the times and that uncovering the root word is more economical than covering it at least 1/2 of the times. We have also seen that our mathematical framework could be used directly to explain some Greenbergian universals. Previous mathematical approaches to word order based on Euclidean distance minimization [8, 10] have not addressed the problem of the ordering of the triplet (S, V, O) and the convering of the root in syntactic dependency trees studied in this article. [8] provides statistical evidence suggesting that distance minimization operates in actual syntactic dependency trees and derives the distribution of the Euclidean distance between syntactically linked words in sentences using both the maximum entropy principle and the assumption that the mean distance between syntactically linked words is constrained to a small value. [8] does not address problems that are relevant to typology of word order universals, at least in the way the discipline is traditionally understood. [10] addresses one of the two conditions of the property of projectivity, i.e. the absence of link crossings in syntactic dependency trees [26]. Here we have studied the other necessary condition for projectivity: the uncovering of the root. This article and [10] provide a complete view of projectivity (though in a simplified way). Our novel findings provide further theoretical support for the hypothesis that the limited resources of the brain introduce biases towards certain word orders [20, 18, 13].

Our model predicts that a priori SVO should be prefered over SOV, but SVO is not the first but the second most frequent word order. This disagreement makes obvious the limits of the model but should no be seen as a big mistake since SVO is the second most frequent word order. Besides, our model predicts that both SVO and OVS should be the most frequent word orders. If SVO and OVS are the most economical a priori, why is OVS rare? This is apparently a big mistake since OVS is disproportionally rare with regard to SVO. In order to address these questions,

18 *Ramon Ferrer-i-Cancho*

we speculate on a very simplified version of the adaptative landscape on which evolution of languages relies and put forward a hypothesis for understanding the frequency of the possible six of orderings of S, V and O in world languages. First, from the point of view of economy, there is a tendency towards orders SVO and its symmetric, i.e. OVS, as explained in Section 5.1. We assume that SVO and OVS are two attractors of the word order dynamics of a language when only distance matters, as argued in Section 5.1. Second, we assume that there is a conflict between SVO and OVS that can be viewed from two perspectives:

- In some languages, the symmetry between the attractors SVO and OVS has not been broken. For this reason, 14% of languages in [5] lack a preference for a specific word order.
- In the remaining languages (the overwhelming majority), the symmetry between SVO and OVS has been broken. Word order counts suggest that the majority has broken the symmetry in favour of SVO and nearby word orders in the permutation ring (Fig. 4) as dominant word orders. For this reason, 35% of languages have SVO as a dominant word order and 96% have an order that is either SVO or its nearest neighbours in the permutation ring (Fig. 4). In order to provide support for the hypothesis that there is a tendency towards SVO (once the symmetry is broken), we consider the correlation between the number of languages that have a certain word order as dominant and the distance in edges between SVO and the remaining word orders in the permutation ring (for instance, the distance between SVO and itself is 0, the distance between SVO and OSV is two). Spearman's correlation lower-tailed test gives $\rho = -0.794$ with $p = 0.03$, indicating that languages grow in number as SVO is approached. This finding can also be interpreted as evidence of a tendency or languages to move away from OVS (notice that the distance to OVS is a linear function of the distance to OVS). Notice that that a symmetry breaking is a null hypothesis for the empirical preference (Table 2) for SVO and nearby word orders in the permutation ring. Alternative hypotheses would be based on assuming that orders near SVO in the permutation ring have been preferred because they have some sort of advantage over OVS and its neighbours in the permutation ring.
- Languages that have a dominant word order different than the attractors suggest that Euclidean distance minimization is not only factor determining word order. However, if such distance minimization actually exists and is strong enough, these conflicting languages must have adopted word order rules to fight against the a priori advantage of the attractors. In particular, the fact that that SOV is the most frequent word order is not a sign that Euclidean distance minimization fails to operate. As we have seen, Greenberg's Universal 16 and 17a can be explained as an adaptation of SOV and VSO to reduce the a priori advantage of SVO.

Here we have made a simplified approach to the S, V O ordering problem because we have assumed that the blocks S, V and O are the only components of the sentence. A priori, we do not expect that this happens very often in real sentences because actual sencences may include other components (e.g., a time complement). However, one could use a large corpus to extract a sufficiently large samples of pure sentences where only S, V and O are present. If this three components, are always located in the boundaries of a real sentence (e.g., adjuncts) then our approach may still be approximately valid. If the components fall between any pair made of S, V and O then our framework is no longer valid. From an empirical point of view, one could collect a sufficiently large list of sentence where the constituents S, V and O (in any possible ordering) appear consecutively (without interruptions from other constituents) in large corpora. We have also made a simplified approach to the covering-of-the-root problem. We have assumed that there is only one syntactic link that could cover the root word. In contrast, our theoretical framework for studying the covering of a vertex, is not only valid for a covered root word but also valid for any kind of potentially covered vertex. Therefore, the projectivity constraint may be too specific when requiring the uncovering of the root word only and forgetting about the roots of subtrees. A bias for uncovering a vertex is a general pressure to the light of our model. However, this pressure may be weaker in coverings involving short distances than in coverings involving long distances (there is less to lose in short links) but we have seen that a bias for uncovering the root decreases as the length of the sentence increases. For this reasons, the distribution of covering arc lengths should be investigated in future studies.

We mentioned above that the goodness of our idealization depends on the successful predictions it can make. We have sucessfully tested our model with some Greenbergian universals but more universal biases should be studied. Our theoretical framework could be the basis for future psycholinguistics or corpus linguistics studies. For instance, one could study the proportion of times that SVO was chosen when it was the best option according to our equations and also the proportion of times that SVO was not chosen and it was not the best option according to our equations. Something similar could be done concerning the the covering problem.

Acknowledgments

20 *Ramon Ferrer-i-Cancho*

environment, unless a written consent of all the persons in the world is obtained.

References

[1] B. Bollobás. *Modern graph theory*. Graduate Texts in Mathematics. Springer, New York, 1998.

[2] W. J. Conover. *Practical nonparemetric statistics*. Wiley, New York, 1999. 3rd edition.

[3] F. de Saussure. *Cours de linguistic générale*. Payot, Paris, 1916.

[4] J. Díaz, J. Petit, and M. Serna. A survey of graph layout problems. *ACM Computing surveys*, 34:313–356, 2002.

[5] M. Dryer. Order of subject, object and verb. In M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie, editors, *The world atlas of language structures*, chapter 81. Oxford University Press, Oxford, 2005.

[6] M. S. Dryer. The greenbergian word order correlations. *Language*, 68(1):80–138, 1992.

[7] G. Fenk-Oczlon and A. Fenk. Zipf's tool analogy and word order. *Glottometrics*, 5:22–28, 2002.

[8] R. Ferrer i Cancho. Euclidean distance between syntactically linked words. *Physical Review E*, 70:056135, 2004.

[9] R. Ferrer i Cancho. Zipf's law from a communicative phase transition. *European Physical Journal B*, 47:449–457, 2005.

[10] R. Ferrer i Cancho. Why do syntactic links not cross? *Europhysics Letters*, 76(6):1228–1235, 2006.

[11] R. Ferrer i Cancho. On the universality of Zipf's law for word frequencies. In P. Grzybek and R. Köhler, editors, *Exact methods in the study of language and text. To honor Gabriel Altmann*, pages 131–140. Gruyter, Berlin, 2007.

[12] E. Gibson. The dependency locality theory: a distance-based theory of linguistic complexity. In *Image, language, brain.*, pages 95–126. The MIT Press, Cambridge, MA, 2000.

[13] E. Gibson and N. J. Pearlmutter. Constraints on sentence comprehension. *Trends in Cognitive Sciences*, 2:262–268, 1998.

[14] T. Givón. *On understanding grammar*. Academic Press, New York, 1979.

[15] J. Greenberg. Some universals of grammar. In J. Greenberg, editor, *Universals of language*. MIT Press, Cambridge, 1968.

[16] J. H. Greenberg. Some univerals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg, editor, *Universals of language*, pages 73–113. MIT Press, London, 1963.

[17] J. H. Greenberg. *Language Universals: with Special Reference to Feature Hierarchies*. Mouton, 1966.

[18] D. Grodner and E. Gibson. Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29:261–290, 2005.

[19] M. Haspelmath, M. S. Dryer, D. Gil, and B. Comrie, editors. *The world atlas of language structures*. Oxford University Press, Oxford, 2005.

[20] J. A. Hawkins. *A performance theory of order and constituency*. Cambridge University Press, New York, 1994.

[21] R. Hudson. *Word Grammar*. Blackwell, Oxford, 1984.

[22] S. Kirby. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C. K. *et al.*, editor, *The evolutionary emergence of language*, pages 303–322. Cambridge University Press, Cambridge, 2000.

[23] S. Kirby. Learning, bottlenecks and the evolution of recursive syntax. In T. Briscoe, editor, *Linguistic evolution through language acquisition*, pages 173–204. Cambridge

University Press, Cambridge, 2002.

[24] W. J. M. Levelt, R. B. L. Page, and H. C. Longuet-Higgins. The speaker's linearization problem. *Journal Philosophical Transactions of the Royal Society of London, Series B*, 295(1077):305–315, 1981.

[25] C. N. Li, editor. *Subject and Topic.* Academic Press, New York, 1976.

[26] I. Mel'čuk. *Dependency syntax: theory and practice.* State of New York University Press, Albany, 1988.

[27] I. Mel'čuk. Language: Dependency. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social and Behavioral Sciences*, pages 8336–8344. Pergamon, Oxford, 2002.

[28] D. Sleator and D. Temperley. Parsing English with a link grammar. Technical report, Carnegie Mellon University, 1991.

[29] R. Vulanović. The combinatorics of cases and word order. *Research on language computation*, 3:107–129, 2005.

[30] R. Vulanović. Grammar efficiency and complexity. *Grammars*, 6:127–144, 2005.

[31] G. K. Zipf. *Human behaviour and the principle of least effort. An introduction to human ecology.* Hafner reprint, New York, 1972. 1st edition: Cambridge, MA: Addison-Wesley, 1949.

Appendix

Here we aim to use recent data [19] to study if word orders have preferences for placing the noun after the adjective (AN) or the adjective after the noun (NA). Combining the information about word order and the placement of the adjective, the data in [19] yields Table 6. From this table, six 2×2 contingency tables (one for each ordering of SVO) can be obtained (Table 7). For constructing these contingency tables, languages lacking a dominant word order are exlucded because we aim to study the relationship between the placement of adjective and dominant word orders. Fisher's exact test [2] can be applied to each of these contingency tables to find out if a certain word order has a preference for the placement of adjectives. From this analysis, the only statistically significant associations at a significance level of 0.05 are SOV with AN and SVO with NA (p-value$< 10^{-5}$ in both cases). Therefore, Greenberg's Universal 17, stating that "with overwhelmingly more than chance frequency, languages with dominant order VSO have the adjective after the noun [17]" is no longer supported by recent data. Interestingly, we have found two statistically significant tendencies that were not reported in Greenberg's original studies. We propose to introduce a variant of Universal 17, i.e. Universal 17a, stating that

"With overwhelmingly more than chance frequency, languages with dominant order SVO have the adjective after the noun and languages with dominant order SOV have the adjective before the noun".

22   *Ramon Ferrer-i-Cancho*

	NA	AN	No dominant order	Only internally headed
OSV	3	0	0	0
OVS	4	2	0	0
SOV	223	166	19	2
SVO	303	56	24	0
VOS	14	7	1	0
VSO	56	16	7	0
No dominant order	65	46	24	1

Table 6. Ordering of S, V and O versus the placement of the adjective of a noun (NA: Noun-Adjective; AN: Adjective-Noun). Inner cells contain the number of languages with a certain row and column feature. Data obtained from [19].

	NA	AN
OSV	3	0
non-OSV	600	247
OVS	4	2
non-OVS	599	245
SOV	223	166
non-SOV	380	81
SVO	303	56
non-SVO	300	191
VOS	14	7
non-VOS	589	240
VSO	56	16
non-VSO	547	231

Table 7. Ordering of S, V and O versus the placement of the adjective of a noun (NA: Noun-Adjective; AN: Adjective-Noun). Inner cells contain the number of languages with a certain row and column feature. Data derived from Table 6.