Why do syntactic links not cross?

R. Ferrer i Cancho[1]

[1] *Departament de Física Fonamental, Universitat de Barcelona. Martí i Franquès 1, 08028 Barcelona, Spain.*

PACS. 89.75.Hc – Networks and genealogical trees.
PACS. 87.53.Wz – Monte Carlo applications.
PACS. 89.90.+n – Other topics in areas of applied and interdisciplinary physics.

Abstract. – Here we study the arrangement of vertices of trees in a 1-dimensional Euclidean space when the Euclidean distance between linked vertices is minimized. We conclude that links are unlikely to cross when drawn over the vertex sequence. This finding suggests that the uncommonness of crossings in the trees specifying the syntactic of sentence could be a side-effect of minimizing the Euclidean distance between syntactically related words. As far as we know, nobody has provided a successful explanation of such a surprisingly universal feature of languages that was discovered in the 60s of the past century by Hays and Lecerf. On the one hand, support for the role of distance minimization in avoiding edge crossings comes from statistical studies showing that the Euclidean distance between syntactically linked words of real sentences is minimized or constrained to a small value. On the other hand, that distance is considered a measure of the cost of syntactic relationships in various frameworks. By cost, we mean the amount of computational resources needed by the brain. The absence of crossings in syntactic trees may be universal just because all human brains have limited resources.

The interplay between the structure of a network and the physical placement of vertices in a Euclidean space has been the subject of various studies (e.g., [1–3]). [1] proposes a way of embedding a scale-free network in a d-dimensional space. [3] introduces a network optimization model generating various kinds of trees (e.g., minimum spanning trees and shortest path trees) in a 2-dimensional space. [2] studies the Euclidean distance in trees where vertices are words and the network defines the syntactic structure of a sentence of the kind shown in Fig. 1 (a). The structure of the sentence has been specified using the dependency grammar formalism [4, 5]. Links indicate syntactic dependencies between words. Link direction is not relevant here because we are only concerned about the Euclidean distance between linked words. The reader interested in more details about this formalism and further examples should refer to [4, 5] or [2]. The motivation of this article comes from these syntactic trees although the results we will present here can be extended to other kinds of networks embedded in 1-dimensional spaces. All the studies summarized above consider that minimizing the physical distance between vertices is a key factor of the shape of the network. We will show that this factor can explain a property of syntactic trees that has remained unexplained for decades since its discovery in the early 60s of the past century by Hays and Lecerf [4]: arcs

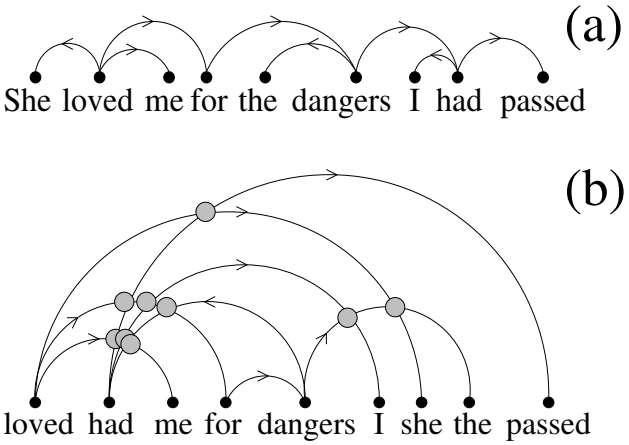

Fig. 1 – (a). A sentence and its syntactic structure. Arcs follow the conventions in [4]. Here vertices are words and the arcs stand for syntactic dependencies. According to [4], arcs go from a head to its modifier. The pronoun 'she' and the verb 'loved' are syntactically dependent in the sentence. 'She' is the modifier of the verbal form 'loved', which is its head ('she' is the subject of 'loved). Similarly, the action of 'loved' is modified by 'me' ('me is the object of the verbal form 'loved'). (b). The structure of the sentence in (b) is the same as that of A but the sequence of vertices is a random permutation of that of (a). Links can only be drawn on the half plane formed by the straight line passing through the row of words. Gray circles indicate edge crossings. While there are no crossings in (a), multiple crossings can be seen in (b).

do not usually cross when drawn over words. Fig. 1 (a) and (b) show a tree without and with crossings, respectively (gray circles indicate crossings). Let us define crossings formally. Suppose that $\pi(v)$ is the position of word v in a sentence. In Fig. 1 (a) we have $\pi(she) = 1$, $\pi(loved) = 2$, $\pi(me) = 3$ and so on. Here we take a specific definition of crossing [4, 6]: the arcs of two pair of words (u, v) and (x, y) such that $\pi(u) < \pi(v)$ and $\pi(x) < \pi(y)$ cross if (and only if) $\pi(u) < \pi(x) < \pi(v) < \pi(y)$ or $\pi(x) < \pi(u) < \pi(y) < \pi(v)$. If the vertices of the sentence in Fig. 1 (a) are scrambled then multiple arc crossings appear (Fig. 1 (b)). Fig. 1 (b) is far from the typical appearance of sentence structures. The majority of sentences in world languages lack crossings with the exception of particular cases [4, 7]. The origins of this universal property constitutes a longstanding problem and is the subject of the present article.

A naive approach to explaining the absence of crossing could be postulating the following principle: the amount of crossings in syntactic structures is minimized. That would trivially explain why arcs do not cross. But that principle would trigger various questions: where does the principle come from? Would it be communicatively advantageous to minimize the amount of crossings? As far as we know, the amount of crossings seems to have no influence on the theoretical amount of information carried by a sentence. In contrast, we will argue that a dramatic reduction of the amount of crossings could be a side-effect of minimizing the distance (in words) between syntactically linked words. The remainder of the article reviews the support that Euclidean distance minimization has in real setences and explains how this principle could be responsible for the exceptionality of edge crossings in syntactic trees.

Evidence about link distance minimization in real sentences comes from three sources: statistics on real sentences, findings in cognitive science and the importance of least effort

principles in language. Statistical evidence that the distance between syntactically linked words is minimized comes from multiple perspectives. First, more than 50% of the links in sentences are formed between consecutive words [2]. This is in great contrast with the a priori probability τ that two linked vertices u and v in a tree are consecutive in the vertex sequence. If n is the length of the sentence in words and we assume that $\pi(u) < \pi(v)$ we have that $\tau = 2/(n-1)$ (if $\pi(u) > 1$ and $\pi(v) < n$; $\tau = 1/(n-1)$ otherwise). It is clear that τ vanishes with n while more than $1/2$ of the links of real sentences are formed between consecutive words. Second, minimizing the sum of the distances between linked vertices on a network where vertices follow a sequence is known as the minimum linear arrangement (m.l.a.) problem [8]. Suppose that we have a network whose set of vertices is V and its set of arcs is A (a directed graph). Suppose that $\pi(v)$ is the position of vertex v . Then, $d(u,v) = |\pi(u) - \pi(v)|$ is the Euclidean distance between vertices u and v (where $u, v \in V$) and $|x|$ is the absolute value of x. For instance, the distance between 'she' and 'loved' in Fig. 1 (a) is 1 while the distance between 'loved' and 'for' is 2. The m.l.a. problem consists of finding the π such that $\Omega(\pi, A) = \sum_{(u,v)\in A} d(u,v)$ is minimum. The value of $\langle d \rangle = \Omega(\pi, A)/n$ of real sentences and the value obtained from a m.l.a. (where vertices are words and links are the syntactic dependencies of a sentence) are similar in order of magnitude but far from the value obtained when assuming that linked words have total freedom for taking distances [2]. From now on, the term distance, when used for syntactically linked words, will refer to the Euclidean distance in words. Third, maximizing a certain entropy while $\langle d \rangle$, the mean arc length in a generic syntactic dependency structure, is constrained, predicts an exponential distribution for the distance between syntactically linked words that is consistent with the real distribution [2].

It is known in cognitive science that the distance between syntactically related items is a measure of the amount of computational resources needed by a sentence [9]. In particular, the distance in words between a word and its syntactic dependents is a reliable measure of cost [10, 11]. By cost, we mean the amount of computational resources needed by the brain. Most of the research about distance-based cost concerns sentence syntactic processing but it has been argued that production undergoes similar constraints [10].

Cost constraints, or equivalently, least effort principles are crucial factors for explaining other universals in quantitative linguistics. For instance, Zipf's law for word frequencies [12] can be explained by maximizing the information transfer of a communication system while the cost of word use is reduced as much as possible [13,14] or by maximizing a certain entropy while word ambiguity is constrained [15].

Here we consider three different arguments for supporting the hypothesis that the exceptionality of edge crossings is a side-effect of link distance minimization. First, a mathematical proof showing that moving a leaf (i.e. a vertex of degree one) right after (or before) its adjacent vertex, will reduce the number of links, if the arc is involved in any crossing (see Appendix). Second, a theoretical argument based on computer simulations on random undirected trees showing that a m.l.a. of vertices reduces dramatically the amount of edge crossings. Those random trees are generated using the following procedure:

1. Start with a network with n vertices and no edges.
2. Choose a pair of vertices at random (all vertices have the same probability to be chosen).
3. Link them if the pair is not yet linked and the network remains without cycles.
4. If the network has less than $n-1$ links go to Step 2.
5. End.

Finding the m.l.a on a generic graph is a very hard computational task. The m.l.a. belongs to the NP-hard class of optimization problems [8,16]. If the sentence structure is a tree, exact computationally affordable algorithms exist [17,18]. A fast heuristic algorithm for solving the m.l.a. problem [19] is used for simplicity here as in [2]. We will generate undirected trees at random and then will compare the initial amount of crossing with the amount of crossings after applying the m.l.a. algorithm. Third, an experimental argument based on computer simulations on a collection of Romanian sentences whose syntactic structure has been specified using the dependency grammar formalism. The argument is the same as the previous one but replacing random trees by real trees from a corpus, i.e. a collection of sentences, where the vertex sequence has been scrambled. The corpus has already been used in previous studies [2, 20]. See the previous references and *http : //phobos.cs.unibuc.ro/roric/DGA/dga.html* for further details about the corpus.

None of the three arguments alone should be taken as solid support for our hypothesis but rather the combination of all three at the same time. Each of the three arguments has its own strong and weak points. The partial theoretical argument covers only a special case but should not be seen as anecdotical. A linear tree where every vertex is linked to the vertex next to it (except for the last vertex), is the worst case situation: only two vertices out of n are leaves. Interestingly, real sentence structures are rarely linear trees (recall Fig. 1). In fact, about one half (46%) of links in the Romanian dependency corpus are formed with a leaf. The theoretical argument based on random trees is intended at covering the general case. As explained in the Appendix, a proof of the general case is not trivial and may turn out to be really hard. The solution adopted here consists of leaving the proof of the general case for future work and use computer simulation as an alternative track. Our solution is not new. Many non-trivial problems in physics are not exactly soluble or do not allow approximated solutions (see for instance [21]). In contrast, computer simulations provide exact results or approximate solutions in problems where mathematical analysis would fail.

The experimental argument has a serious drawback: results are strictly valid only for the corpus under consideration and the conventions adopted for representing the structure of each sentence. Changing the representational conventions, the corpus or the language may lead to different results, although one may argue that this is unlikely. This is important because the exceptionality of crossings seems to be a linguistic universal and we are aimed at showing why this may be so. The virtue of the computer simulations on random trees is skipping the tedious work of testing the hypothesis in as many languages and corpora as possible by abstracting from details such as the language under consideration, the corpus, the age of the speaker/writer, the audience, the topic, etc. For this reason, we must rely on minimal assumptions about what the structure of a sentence is. We assume that sentences are trees (i.e. acyclic connected graphs [22]) and that links are formed by choosing pairs of words at random (a possible reason for not choosing them at random could be trying to mimic the statistical properties of a set of real sentence structures, which would prevent us from achieving our ultimate goal, which is explaining the apparent universality of the exceptionality of crossings).

Since we are interested in the scope of the mathematical proof of a particular case, we will also consider an intermediate situation between random arrangements and minimum linear arrangements: arrangement consisting of moving, each of the leaves after or before its adjacent vertex (our convention is choosing the placement involving the shortest displacement). This configuration will be called a partial m.l.a.

$\langle d \rangle$ versus n (the sentence length in words) has been already been studied in depth [2]. What is new here is $\langle d \rangle$ for partial m.l.a.'s. It can be seen that partial m.l.a.'s give values of $\langle d \rangle$ between scrambled vertex sequence and minimum linear arrangements for random trees (Fig. 2 (a)) and Romanian syntactic dependency trees (Fig. 2 (b)). Moreover, Fig. 2 (b) shows

that real trees have values of $\langle d \rangle$ that are slightly above those of m.l.a's but very different from those of scrambled vertex sequences or partial m.l.a.'s. Fig. 2 (c) and (d) show a drastic reduction of edge crossings after applying the m.l.a. to random trees and scrambled Romanian syntactic dependency trees. Despite thir simplicity, the partial m.l.a. behaves almost like a full m.l.a. for small n. We define C as the number of edge crossings of a tree. In random trees, the m.l.a. gives, on average, $C = 0$ for $n \leq 33$ and $C < 0.02$ for $n > 33$ in the interval of n explored. In Romanian syntactic dependency trees, the m.l.a. gives, on average, $C < 0.24$ for $n \leq 28$ and $C < 1.58$ for $n > 28$. This means that less than one crossing is expected on average for a wide range of n. No tree had crossings in the Romanian corpus used here. Our results suggest that sentences with one or more crossings should be rare, which is consistent with the exceptional nature of crossings in real sentences. In sum, random trees or scrambled real trees have a large amount of crossings that becomes insignificant after applying the m.l.a. algorithm.

The results in Fig. 2 indicate that the absence of crossings in real sentences may be a side effect of minimizing the Euclidean distance between syntactically linked words. As said above, the Euclidean distance is a measure of the cost of a syntactic relationship. This cost is a consequence of the limited computational resources of the human brain [9, 10]. We have found that the ordering of words in real sentences is very similar to that of a m.l.a. but we do not mean that the brain actually performs a m.l.a. when generating sentences. We leave the question of whether human actually performs an m.l.a or not for future research. We do not mean that distance minimization is strictly the only reason for the absence of crossings, but probably the most important factor. In sum, it would be surprising that arcs crossed frequently in world languages, and even more surprising that a sentence contained many crossings, given the high computational demands that crossings indirectly impose.

* * *

We are specially grateful to P. Fernández for technical assistance and making the author aware of the minimum linear arrangement problem. We thank Q. Trullàs, D. Hudson and J. Díaz for helpful comments. This work was funded by the a grant of the Generalitat de Catalunya (FI/2000-00393) and a Juan de la Cierva contract from the Spanish Ministry of Education and Science under the project BFM2003-08258-C02-02.

Appendix

We study the effect of reducing the length of an edge (u, v) on the number of crossings from the theoretical point of view. We assume that the network is a tree. We will use $\pi(v)$ for designating the initial position of v and $\pi'(v)$ for the new position of vertex v once the length of a certain edge has been decreased. We define $\delta(v)$ as the degree of vertex v (i.e. the number of links of v in A) [22]. For simplicity, we consider a particular case, where the target edge is (u, v), where v is a leaf and u is not (i.e. $\delta(u) \geq 2$ and $\delta(v) = 1$). We focus on the case $\pi(u) < \pi(v)$. The treatment of the case $\pi(u) > \pi(v)$ is analogous.

We define $C(u, v)$ as the number of crossings induced by the edge (u, v). We study the change in the number of crossings after moving v from $\pi(v)$ to a new position η, such that $\pi(u) < \eta \leq \pi(v)$. We denote the new value of $C(u, v)$ once v has been moved by $C'(u, v)$. After moving v, the position of some vertices must be recalculated. We adopt the following rearrangement convention: $\pi'(x) = \eta$ if $x = v$, $\pi'(x) = \pi(x)$ if $\pi(x) < \eta$ and $\pi'(x) = \pi(x)+1$ if

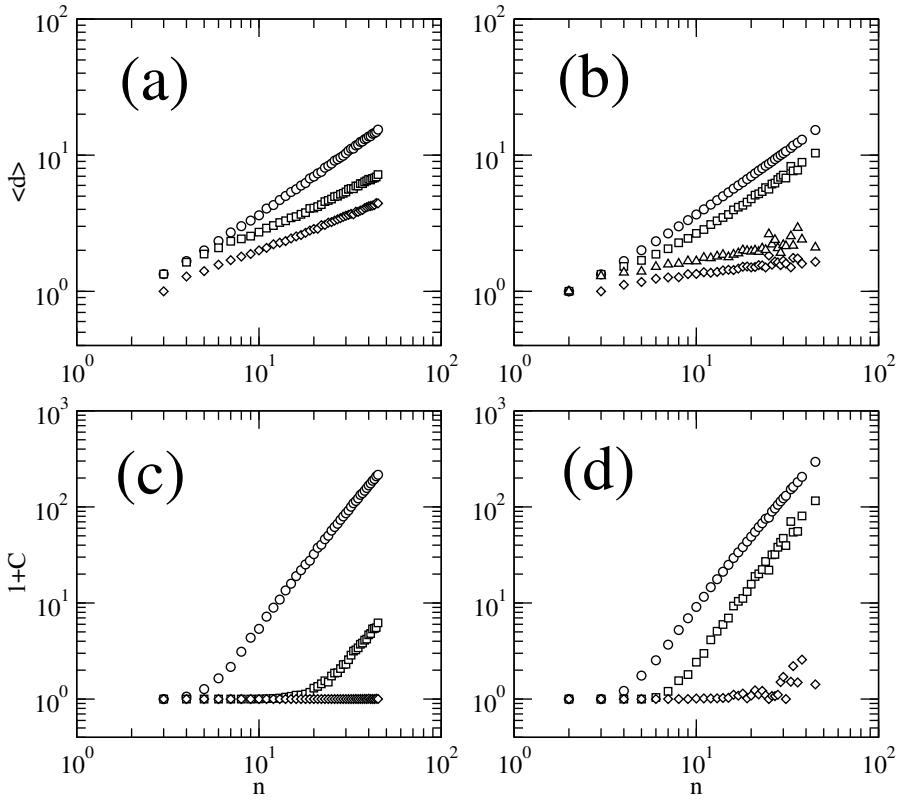

Fig. 2 – The average value of C and $\langle d \rangle$ versus n, the sentence length in words in a collection of trees. C is the number of arc crossings and $\langle d \rangle$ is the mean Euclidean distance between syntactically linked words of a tree whose vertices form a sequence. The series for C have been conveniently rescaled by adding 1 to C. (a) Average $\langle d \rangle$ for random trees. Circles, squares and diamonds identify, respectively, the series for random trees, partial m.l.a. and full m.l.a. (b) Average $\langle d \rangle$ for Romanian syntactic dependency trees. Circles, squares and diamonds stand for, respectively, the series for scrambled real trees, partial m.l.a. and full m.l.a. of real trees. Triangles indicate the average value of $\langle d \rangle$ for trees where words are arranged in the same order as in the original sentences. (c) Average C for random trees. The shapes of each series follow the same conventions of (a). (d) Average C for Romanian syntactic dependency trees. The shapes of each series follow the same conventions of (b). None of the Romanian syntactic dependency trees had crossings, so the corresponding series has been omitted for clarity. In (a) and (c), the averages for random trees (circles) were obtained generating 500 replicas for each value of n. As for real syntactic dependency trees ((b) and (d)), averages for scrambled real trees where calculated over the mean value of 500 scramblings on each real tree. Error bars are not shown in any of the subfigures for the sake of clarity. In general, error bar lengths for the value of C of m.l.a.'s are at most of the same order of the symbol size both in (c) and (d).

$\pi(x) \geq \eta$. We have $C'(u,v) = C(u,v) + \Delta(u,v)$ where $\Delta(u,v)$ is the change in the number of crossings generated by the new vertex positions. We define $\zeta(u,v)$ as the number of different edges that cross the vertices u and v (u and v are not necessarily linked), i.e. edges of the kind (x,y) where $\pi(u) < \pi(x) < \pi(v)$ and $\pi(y) < \pi(u)$ or $\pi(y) > \pi(v)$. We also define $\epsilon(u,v,\eta)$ as the number of different edges that will induce a new crossing with the link (u,v) once v has

been place in position η, i.e. edges of the kind (x, y) such that $\pi(u) < \pi'(x) < \eta < \pi'(y) \leq \pi(v)$ (or $\pi(u) < \pi(x) < \eta \leq \pi(y) < \pi(v)$). It can be easily shown that

$$\Delta(u, v) = \epsilon(u, v, \eta) - \zeta(\pi^{-1}(\eta), v), \tag{A.1}$$

where π^{-1} is the inverse function of π and, $\epsilon(u, v, \eta)$ and $\zeta(\pi^{-1}(\eta), v)$ are, respectively, the amount of crossings that have appeared or disappeared after having shortened (u, v) by moving v towards u. In general, we cannot say if $\Delta(u, v) < 0$ or not (the answer depends on η). If $\eta = \pi(u) + 1$ then $\epsilon(u, v, \eta) = 0$ and thus $\Delta(u, v) \leq 0$ because $\zeta(\pi^{-1}(\pi(u) + 1), v) \geq 0$. In this case, we get $\Delta(u, v) < 0$ if $\zeta(\pi^{-1}(\eta), v) > 0$, that is, if (u, v) was involved in at least one crossing. A proof that shortening an edge decreases the number of crossings (if that edge is involved in crossings) is apparently difficult in a general case (i.e. $\pi(u) < \eta < \pi(v)$ and $\delta(v) \geq 1$). Apparently, simple arguments are only available under restrictions. Eq. A.1 is only valid when $\delta(v) = 1$. If $\delta(v) > 1$, one has to worry about the crossings induced by edges of v that are not formed with u. Shortening an edge may imply lengthening others. Complicated configurations arise when $\delta(v) \geq 2$: pulling v towards u increases the length of the edges formed with vertices that are placed after v ($\delta(v) - 1$ edges in the worst case) and decreases the length of the edges formed with vertices placed before u.

REFERENCES

[1] A. F. Rozenfeld, R. Cohen, D. ben Avraham, and S. Havlin, Physical Review Letters **89**(21), 218701 (2002).
[2] R. Ferrer i Cancho, Physical Review E **70**, 056135 (2004).
[3] M. Barthélemy and A. Flammini, arXiv:physics/0601203 (2006).
[4] I. Melčuk, *Dependency Syntax: Theory and Practice* (SUNY, Albany, 1988).
[5] R. Hudson, *Word Grammar* (Blackwell, Oxford, 1984).
[6] P. Flajolet and M. Noy, Discrete Mathematics **204**(1-3) (1999).
[7] I. Melčuk, in V. Agel, L. Eichinnger, H.-W. Eroms, P. Hellwig, H. J. Herringer, and H. Lobin, eds., *Dependency and valency. An international handbook of contemporary research* (W. de Gruyter, New York, 2003), pp. 288–229.
[8] J. Díaz, J. Petit, and M. Serna, ACM Computing surveys **34**, 313 (2002).
[9] E. Gibson and N. J. Pearlmutter, Trends in Cognitive Sciences **2**, 262 (1998).
[10] D. Grodner and E. Gibson, Cognitive Science **29**, 261 (2005).
[11] J. A. Hawkins, *A performance theory of order and constituency* (Cambridge University Press, New York, 1994).
[12] 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.
[13] R. Ferrer i Cancho and R. V. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2003).
[14] R. Ferrer i Cancho, Eur. Phys. J. B **47**, 449 (2005).
[15] R. Ferrer i Cancho, Physica A **345**, 275 (2005), doi:10.1016/j.physa.2004.06.158.
[16] M. R. Garey and D. S. Johnson, *Computers and intractability: a guide to the theory of NP-completeness* (W. M. Freeman, San Francisco, 1979).
[17] Y. Shiloach, SIAM J. Comput. **8**(1), 15 (1979).
[18] F. R. K. Chung, Comp. & Mahts. with Appls. **10**(1), 43 (1984).
[19] Y. Koren and D. Harel, in *Proceedings of 28th Inter. Workshop on Graph-Theoretic Concepts in Computer Science (WG'02)* (Springer Verlag, Berlin, 2002), vol. 2573 of *Lecture Notes in Computer Science*, pp. 293–306.
[20] R. Ferrer i Cancho, R. V. Solé, and R. Köhler, Physical Review E **69**, 051915 (2004).
[21] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids* (Claredon, Oxford, 1987).
[22] B. Bollobás, *Modern graph theory*, Graduate Texts in Mathematics (Springer, New York, 1998).