Euclidean distance between syntactically linked words

Ramon Ferrer i Cancho*
ICREA-Complex Systems Laboratory, Universitat Pompeu Fabra, Dr. Aiguader 80, 08003 Barcelona, Spain
and INFM udR Roma 1, Dipartimento di Fisica, Università La Sapienza, Piazzale A. Moro 5, 00185 Roma, Italy
(Received 26 April 2004; published 30 November 2004)

We study the Euclidean distance between syntactically linked words in sentences. The average distance is significantly small and is a very slowly growing function of sentence length. We consider two nonexcluding hypotheses: (a) the average distance is minimized and (b) the average distance is constrained. Support for (a) comes from the significantly small average distance real sentences achieve. The strength of the minimization hypothesis decreases with the length of the sentence. Support for (b) comes from the very slow growth of the average distance versus sentence length. Furthermore, (b) predicts, under ideal conditions, an exponential distribution of the distance between linked words, a trend that can be identified in real sentences.

DOI: 10.1103/PhysRevE.70.056135 PACS number(s): 89.75.−k, 89.20.−a

I. INTRODUCTION

Dependency grammar is a class of grammatical formalisms [1–3] 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). 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. The dependency grammar formalism distinguishes some cases, such as coordination, where there is no clear direction [4].

The statistical structure of global syntactic dependency networks has recently received attention [5]. Those networks have words as nodes. A pair of words is linked if that pair has appeared syntactically connected at least once in a corpus (i.e., collection of sentences).

Here we focus on the Euclidean (or physical) distance between syntactically linked words in sentences. 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 following words. 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 $d(u,v)=d(v,u)$. We are only interested in the distance between connected words. Table I 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 I would remain the same except for $d(she,loved)=8$.

There are reasons for thinking that the distance between syntactically linked words is constrained. The language faculty is constrained in many ways. Lung capacity imposes limits on the length of actual spoken sentences, whereas working memory [6] imposes limits on the complexity of sentences if they are to be understandable [7]. It is reasonable to think that distantly related words pose problems to the brain machinery that has to produce or process a certain sentence. The fact that about 50%–67% of the links in sentences are formed between words at distance 1 and 16%–25% are formed at distance 2 [8] suggests two possibilities: (a) the Euclidean distance between syntactically linked words is minimized or (b) the Euclidean distance between linked words is constrained on average. Various statistical tests indicate that the distance at which syntactic interactions take place is significantly small [8].

The distance between syntactically related items in sentences is a basic ingredient of the cost of a sentence [9,10] and has been used for explaining word order universals [10]. Cost minimization or, in other words, least effort principles are a successful explanation for universals in quantitative linguistics. For instance, Zipf's law [11] for word frequencies can be explained by minimizing hearer and speaker communicative needs [12,13].

The minimization of the topological (or network) distance in complex networks has been studied [14–16]. Minimization of the Euclidean distance has been studied in various topologies: rings [17] and two-dimensional Euclidean spaces [18] (see [19] for more references). Here we focus on a one-dimensional Euclidean space without boundary conditions. An important difference is that we assume that the network structure is fixed. The only freedom is for changing the po-

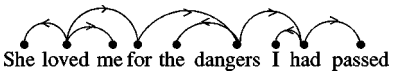

FIG. 1. The syntactic structure of the sentence, "She loved me for the dangers I had passed," following the conventions in [1]. Here vertices are words and the arcs stand for syntactic dependencies. Following the conventions in [1], arcs go from a head to its modifier. The pronoun "she" and the verb "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.

*Electronic address: ramon@pil.phys.uniroma1.it

TABLE I. 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 $[d(word,sender)]$.

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

sitions of words in the sentence, that is changing the function $\pi(v)$.

Section II suggests that the linear arrangement of words in sentences obeys a minimization of the Euclidean distance between words [hypothesis (a)]. Section III suggests that arrangement could be constrained in the mean Euclidean distance between words [hypothesis (b)].

II. EUCLIDEAN DISTANCE MINIMIZATION HYPOTHESIS

Suppose we have a network whose set of vertices is V and its set of arcs is A (a directed graph). Suppose $\pi(v)$ is the position of vertex v. Then, $d(u,v)=|\pi(u)-\pi(v)|$ is the Euclidean distance between vertices u and v (where $u,v \in V$). We are aimed at finding the π such that $\Omega(\pi,A)=\Sigma_{(u,v)\in A}d(u,v)$ is minimum. Minimizing Ω, as defined here, is known as the minimum linear arrangement (m.l.a.) problem [20]. Here we will consider if the Euclidian distance between syntactically related words is minimized. The problem that minimization must solve is exactly the m.l.a. in computer science [20]. Two different sources of data were used for the present study. Both are collections of sentences with its syntactic dependency structure. Both data sets have been already used in [5]. First, a Romanian corpus was formed by all sample sentences in the Dependency Grammar Annotator website [21]. It contains 21 275 words and 2340 sentences. Second, a Czech corpus was used [22,23] having approximately 563 067 words and 31 701 sentences. Many sentence structures are incomplete in the Czech corpus (i.e., they have fewer than $n-1$ links, where n is the length of the sentence in words). The proportion of links provided with regard to the theoretical maximum is about 0.65. When having complete structures was critical, only the Romanian corpus was used. Punctuation marks were absent, so distances between words are true distances in both cases.

We define the average value of d, the distance between linked vertices, as

$$\langle d\rangle = \frac{1}{n-1}\Omega(\pi,A),$$

where n is the length of the sentence in words (notice $|A|=n-1$). Alternatively, we may define $\langle d\rangle$ as

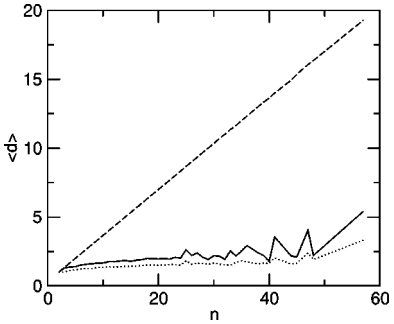

FIG. 2. The average value of $\langle d\rangle$, the mean edge length (in words), versus the length (in words) of the sentence, n, for real sentences (solid line) and the corresponding minimum linear arrangements (dotted line). Here the Romanian corpus is used. A control $\langle d\rangle$ was calculated by scrambling the words in every sentence 1000 times and averaging $\langle d\rangle$ (long dashed). The latter case is $\langle d\rangle=(n+1)/3$, as expected.

$$\langle d\rangle = E[d] = \sum_{d=1} dP(d), \tag{1}$$

where E is the expectation operator and $P(d)$ is the probability that two linked words are at distance d. Figure 2 shows $\langle d\rangle$ as a function of n for real Romanian sentences. The real $\langle d\rangle$ is compared against a null hypothesis (a control series) and the value obtained by a m.l.a. As for the null hypothesis, it is calculated on real sentences by scrambling the position of vertices (while the network structure remains the same) and calculating Ω again [Ω is used instead of $\Omega(\pi,A)$ for brevity]. It follows for the latter case that

$$P(d) = \frac{2(n-d)}{n(n-1)}. \tag{2}$$

Replacing the previous equation into Eq. (1) we get

$$\langle d\rangle = \frac{n+1}{3} \tag{3}$$

after some algebra. Incidentally, Eq. (3) is the same as the average vertex-vertex distance of a linear graph [24]. As for the m.l.a., a fast heuristic algorithm for solving the m.l.a. problem [25] is used for simplicity. Finding the m.l.a on a generic graph is a very hard computational problem [20,26]. If the network is a tree, exact computationally affordable algorithms exist [27,28]. Numerical calculations up to $n=11$ showed that the algorithm in [25] always finds the optimum on trees. Figure 2 shows that real $\langle d\rangle$ is significantly small, given how far the real series is from the upper bound provided by the null hypothesis in Eq. (3). Figure 2 supports the hypothesis that real sentences may minimize Ω to some extent. The fact that $\langle d\rangle$ for real sentences is greater than that of the heuristic approximation shows that using the exact algorithm for trees [27,28] is not necessary in this context.

We define the ratio

$$\Gamma = \Omega_{real}/\Omega_{mla},$$

where Ω_{real} and Ω_{mla} are, respectively, the average value of Ω for the Romanian collection of sentences and that of the

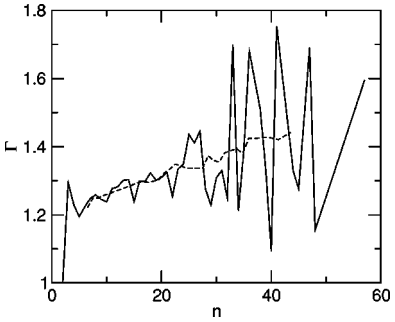

FIG. 3. The optimization ratio Γ versus sentence length in words (solid line) in the Romanian corpus. Running averages show a tendency of Γ to grow with n (dashed line).

corresponding m.l.a.'s. Γ is a growing function of n, the sentence length (Fig. 3). Therefore, the shorter the sentence, the higher the support for the Euclidean distance minimization hypothesis.

III. CONSTRAINED EUCLIDEAN DISTANCE HYPOTHESIS

Figure 2 suggests that $\langle d \rangle$ is constrained in real sentences because $\langle d \rangle$ is a very slowly growing function of n (a linear fitting gives $\langle d \rangle = 1.163 \pm 0.039n$). Additional support for that constraint comes from the computational limitations of the brain for dealing with distant syntactically linked words [9]. Working memory [6] carries on the load of distant linked words. So the hypothesis here is twofold: (a) there is a limited amount of resources (e.g., working memory) for producing and processing sentences that should not be exceeded and (b) $\langle d \rangle$ is a good measure of the amount of resources required by the structure of sentence.

The distance between linked words of an ideal language constraining $\langle d \rangle$ can be calculated. We can predict $P(d)$, the probability that two linked words are at distance d using the maximum entropy principle [29,30] for obtaining $P(d)$ when the arc mean length $\langle d \rangle$ is constrained as Fig. 2 suggests. Thus, we get (see the Appendix)

$$P(d) = a(n-d)e^{-\beta d}, \tag{4}$$

where β is a parameter and

$$a = \left(\sum_{d=1}^{n-1} (n-d)e^{-\beta d} \right)^{-1}.$$

β is a parameter satisfying

$$\langle d \rangle = \sum_{d=1}^{n-1} d(n-d)e^{-\beta d}.$$

For large n (see the Appendix) we have

$$\beta \approx \frac{n - \langle d \rangle \pm (\langle d \rangle^2 - 10n\langle d \rangle - n^2)^{1/2}}{2\langle d \rangle n} \tag{5}$$

and n and $\langle d \rangle$ are the only parameters.

Support for maximizing the entropy of $\{P(d)\}$ comes from the following. The minimum entropy is given by an arrangement where $\langle d \rangle = 1$. The only networks that can achieve such a small distance are linear graphs where all links are formed between consecutive words in the sentence. A linear graph is a connected graph without cycles where all vertices have two connections except two vertices in the extremes which have a single connection [24]. For instance, the sentence "I eat potatoes," whose links are {"I," "eat"} and {"eat," "potatoes"} is a linear graph of that kind. The problem of a linear a graph is that it is a very specific structure. For instance, the graph in Fig. 1 is not a linear graph, so it cannot inherently achieve $\langle d \rangle = 1$. Sentence structure imposes links whose vertices pairs cannot be arranged consecutively. Thus, sentence structure induces maximizing the entropy of $\{P(d)\}$.

IV. DISCUSSION

We have considered two hypotheses concerning the Euclidean distance between syntactically linked words: (a) $\langle d \rangle$ is minimized and (b) $\langle d \rangle$ is constrained while the entropy is maximized. First, we examine the support for hypothesis (a). We have found that real $\langle d \rangle$ is slightly above what a minimum linear arrangement would dictate but very far from the null hypothesis, the expected $\langle d \rangle$ when there is not minimization at all (Fig. 2). Real $\langle d \rangle$ is significantly small. We have also seen that Γ, the ratio between real $\langle d \rangle$ and m.l.a. $\langle d \rangle$, increases with the length of the sentence (Fig. 3). Second, we examine the support for hypothesis (b). If $\langle d \rangle$ was constrained in real sentences in full, a straight line in linear-log scale with the predicted exponent would be expected. Although Eq. (4) is close to the real values for short distances, it cannot directly explain the exponential trend with different slope for long distances (Fig. 4). The slower decrease in $P(d)$ for long distances suggests the presence of factors such as word order rules preventing $P(d)$ from decreasing as fast as a pure Euclidean distance constraint would dictate. Let us illustrate what could be happening at long distances with a simple example (a short phrase is chosen for simplicity). The phrase "beautiful black car," whose edges are {"beautiful," "car"} and {"black," "car"}, gives $\langle d \rangle = 3/2$. A better arrangement would be "beautiful car black," giving $\langle d \rangle = 1$, but that would violate English grammar rules. That type of grammatical conflicts may also explain the growth of Γ with n.

Our study provides support for hypotheses (a) and (b). Both hypotheses are complementary although one could be a consequence of the other. Constraining $\langle d \rangle$ to a certain value is similar to minimizing $\langle d \rangle$ if that value is sufficiently small. In other words, (a) could be, to some extent, a side effect of (b). Distant connections are so expensive in terms of memory that they are very unlikely to happen, but if the sentence structure is complex enough, links between nonconsecutive words cannot be avoided. Distance minimization or constrained distance seems a consequence of limited brain resources.

ACKNOWLEDGMENTS

We thank Pau Fernández for technical assistance concerning the minimum linear arrangement algorithm. Francesca

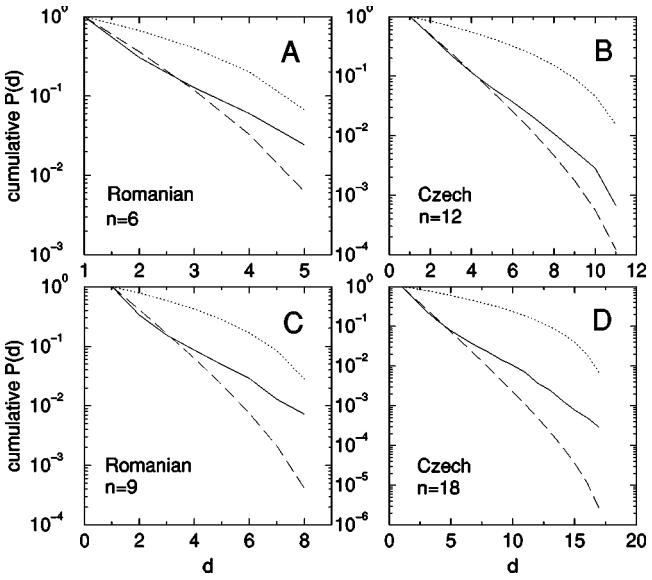

FIG. 4. The cumulative $P(d)$, where $P(d)$ is the probability an arc links words at distance d. Real values (solid lines) can be compared to that of the null hypothesis (dotted lines) and the maxent exponential model (dashed lines). (A) Romanian sentences having the typical length $L^*=6$. (B) Czech sentences having the typical length $L^*=12$. (C) Romanian sentences having the mean length $\langle L\rangle \approx 9$ D. Czech sentences having the mean length $\langle L\rangle \approx 18$. Real $P(d)$ clearly differs from the null hypothesis and approaches a straight line in linear-log scale, agreeing with the exponential prediction derived in this paper for short distances and changing the slope but keeping the exponential trend for long distances.

Colaiori made helpful comments. We are grateful to the Dependency Grammar Annotator people for the Romanian data and Ludmila Uhlířová for the opportunity to analyze the Czech data. This work was supported by the Institució Catalana de Recerca i Estudis Avançats (ICREA), the Grup de Recerca en Informàtica Biomèdica (GRIB), a grant of the Generalitat de Catalunya (No. FI/2000-00393), and the FET Open Project COSIN (No. IST-2001-33555).

APPENDIX: THE MAXIMUM ENTROPY PRINCIPLE

p_d, the probability that two linked words are at distance d [$P(d)$ in the main text], can be derived using the minimum entropy principle [29,30]. Knowing that the prior distribution is $\mathcal{P}_d=2(n-d)/n(n-1)$ and assuming there is no further constraint other than normalization, we may define the functional

$$E = H_B - \alpha \sum_{d=1}^{n-1} p_d,$$

where H_B is the Bayesian entropy defined as

$$H_B = -\sum_{d=1}^{n-1} p_d \ln \frac{p_d}{\mathcal{P}_d}.$$

$\partial E/\partial p_d = 0$ leads to

$$p_d = \mathcal{P}_d e^{-1-\alpha}.$$

The constraint $\Sigma_{d=1}^{n-1} p_d = 1$ gives $p_d = \mathcal{P}_d$ as expected.

Assuming $\langle d\rangle = \Sigma_{d=1}^{n-1} d p_d$, the average distance between linked words, is constrained, we may define the functional

$$E = H_B - \alpha \sum_{d=1}^{n-1} p_d - \beta \sum_{d=1}^{n-1} d p_d.$$

Thus, $\partial E/\partial p_d = 0$ leads to

$$p_d = \mathcal{P}_d e^{-1-\alpha-\beta d},$$

which we may write as

$$p_d = a(n-d)e^{-\beta d},$$

with

$$a = \frac{2e^{-1-\alpha}}{n(n-1)}.$$

The constraint

$$\sum_{d=1}^{n-1} p_d = 1$$

leads to

$$a = \left(\sum_{d=1}^{n-1} (n-d) e^{-\beta d} \right)^{-1}. \tag{A1}$$

The constraint

$$\sum_{d=1}^{n-1} d p_d = \langle d\rangle$$

leads to

$$a = \frac{\langle d\rangle}{\Sigma_{d=1}^{n-1} d(n-d)e^{-\beta d}}. \tag{A2}$$

Minimizing the function

$$F = (ab - \langle d\rangle)^2,$$

with

$$b = \sum_{d=1}^{n-1} d(n-d)e^{-\beta d},$$

we may obtain the value(s) of β. Knowing

$$\int_0^\infty x^n e^{-ax} dx = \frac{\Gamma(n+1)}{a^{n+1}},$$

we may write Eq. (A1) as

$$a \approx \left(\frac{n\Gamma(1)}{\beta^2} - \frac{\Gamma(2)}{\beta^3} \right)^{-1} \tag{A3}$$

and Eq. (A2) as

$$a \approx \frac{\langle d \rangle}{\frac{n\Gamma(2)}{\beta^2} - \frac{\Gamma(3)}{\beta^3}} \tag{A4}$$

for large n. The right-hand sides of Eqs. (A3) and (A4) together give

$$\langle d \rangle n\beta^2 - (\langle d \rangle + n)\beta + 2 \approx 0. \tag{A5}$$

Therefore, we have

$$\beta \approx \frac{n - \langle d \rangle \pm (\langle d \rangle^2 - 6n\langle d \rangle + n^2)^{1/2}}{2\langle d \rangle n}.$$

[1] I. Melčuk, *Dependency Syntax: Theory and Practice* (SUNY, Albany, 1988).
[2] R. Hudson, *Word Grammar* (Blackwell, Oxford, 1984).
[3] D. Sleator and D. Temperley, Tech. Rep., Carnegie Mellon University (1991).
[4] I. Melčuk, in *International Encyclopedia of the Social and Behavioral Sciences*, edited by N. J. Smelser and P. B. Baltes (Pergamon, Oxford, 2002), pp. 8336–8344.
[5] R. Ferrer i Cancho, R. V. Solé, and R. Köhler, Phys. Rev. E **69**, 051915 (2004).
[6] A. Baddeley, *Working Memory* (Clarendon, Oxford, 1986).
[7] M. D. Hauser, N. Chomsky, and W. T. Fitch, Science **298**, 1569 (2002).
[8] R. Ferrer i Cancho (unpublished).
[9] E. Gibson, *Image, Language, Brain* (MIT Press, Cambridge, MA, 2000), pp. 95–126.
[10] J. A. Hawkins, *A Performance Theory of Order and Constituency* (Cambridge University Press, New York, 1994).
[11] G. K. Zipf, *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Hafner, New York, 1972).
[12] R. Ferrer i Cancho and R. V. Solé, Proc. Natl. Acad. Sci. U.S.A. **100**, 788 (2003).
[13] R. Ferrer i Cancho, Physica A (to be published). (to be published).
[14] R. Ferrer i Cancho and R. V. Solé, in *Statistical Physics of Complex Networks*, Lecture Notes in Physics (Springer, Berlin, 2003).
[15] V. Venkatasubramanian, S. Katare, P. R. Patkar, and F.-P. Mu, Comput. Chem. Eng. (to be published).
[16] T. Nishikawa, A. E. Motter, Y.-C. Lai, and F. C. Hoppensteadt, Phys. Rev. E **66**, 046139 (2002).
[17] N. Mathias and V. Gopal, Phys. Rev. E **63**, 021117 (2001).
[18] A. K. S. S. Manna, J. Phys. A **36**, L279 (2003).
[19] M. Barthélemy, Europhys. Lett. **63**, 915 (2003).
[20] J. Díaz, J. Petit, and M. Serna, ACM Comput. Surv. **34**, 313 (2002).
[21] The Dependency Grammar Annotator website is http://phobos.cs.unibuc.ro/roric/DGA/dga.html
[22] L. Uhlírova, I. Nebeská, and J. Králík, in *COLING 82, Proceedings of the Ninth International Conference on Computational Linguistics, Prague, 1982*, edited by J. Horecký (North-Holland, Amsterdam, 1982), pp. 391–396.
[23] M. Těšitelová, Academia Praha 249s (1985).
[24] R. Ferrer i Cancho and R. V. Solé, *Optimization in Complex Networks*, Lecture Notes in Physics Vol. 625 (Springer, Berlin, 2003), pp. 114–125.
[25] Y. Koren and D. Harel, in *Proceedings of 28th International Workshop on Graph-Theoretic Concepts in Computer Science (WG'02)*, Lecture Notes in Computer Science Vol. 2573 (Springer Verlag, Berlin, 2002), pp. 293–306.
[26] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, San Francisco, 1979).
[27] Y. Shiloach, SIAM J. Comput. **8**, 15 (1979).
[28] F. R. K. Chung, Comput. Math. Appl. **10**, 43 (1984).
[29] J. N. Kapur, *Maximum Entropy Models in Science and Engineering* (Wiley, New Delhi, 1989), pp. 30–43.
[30] E. W. Montroll and M. F. Shlesinger, J. Stat. Phys. **32**, 209 (1983).