# The structure of syntactic dependency networks: insights from recent advances in network theory.

*Ramon Ferrer i Cancho*[1]

**Abstract.** Complex networks have received substantial attention from physics recently. Here we review from a physics perspective the different linguistic networks that have been studied. We focus on syntactic dependency networks and summarize some recent strong results that suggest new possible ways of understanding the universal properties of world languages.

**In press in "Problems of quantitative linguistics". Edited by Victor Levickij and Gabriel Altmann.**

## INTRODUCTION

In many systems, elements connect and form a network. Examples are words in the syntactic dependency structure of a sentence, species in an ecosystem or web pages in the World Wide Web. In the first case, connections are syntactic dependencies ($x$ is the head of the modifier $y$), in the second case connections are predation relationships (a species $x$ predates on species $y$) and in the third case connections are mouse clicks[2] (a web page $x$ takes to web page $y$ in one mouse click).

In network theory, $x$ and $y$ are called vertices and their connection is called a link, edge or arc. Networks have recently been in the spotlight of physics. The characterization of the statistical properties of real networks and their modelling has been the subject of a considerable amount of research (Albert & Barabási, 2002; Dorogovtsev & Mendes, 2002; Newman, 2003).

Different kinds of linguistic networks have been studied in the literature from a physics perspective: thesaurus networks (Steyvers & Tenenbaum, 2005,2001; Newman, 2003; Albert & Barabási, 2002; Motter *et al.*, 2002; Kinouchi *et al.*, 2002), WordNet (Steyvers & Tenenbaum, 2005,2001; Sigman & Cecchi; 2002), word association networks (Steyvers & Tenenbaum, 2005,2001; Capocci *et al.*, 2005), word co-occurrence networks (Ferrer i Cancho & Solé, 2001;

[1] Address for correspondence: Ramon Ferrer i Cancho, Dip. di Fisica, Università 'La Sapienza', Piazzale A. Moro 5, ROMA 00185, ITALY. E-mail: ramon.ferrericancho@gmail.com

[2] To be more precise, connections are HTML links.

Dorogovtsev & Mendes, 2001, 2003a; Milo *et al.*, 2004) and syntactic dependency networks (Ferrer i Cancho *et al.*, 2004; Ferrer i Cancho, 2004, 2005d; Ferrer i Cancho *et al*., 2005b). The next section contains an overview of these linguistics networks. The structure of syntactic dependency networks is explored in more detail in two further sections that summarize the latest advances.

## A QUICK OVERVIEW OF LINGUISTIC NETWORKS

Thesauri are formed by lists of entries where the first word of the entry is the root word. The words after the root are roughly synonymous words (Table 1). All vertices are words in the network. In Steyvers & Tenenbaum (2005, 2001), two words are linked if one word has been the root word of the other in at least one entry of the thesaurus. Essentially, two kinds of thesaurus have been studied: based on Roget's thesaurus (Steyvers & Tenenbaum, 2005,2001; Motter *et al.*, 2002; Newman, 2003) and based on Merrian-Webster's thesaurus (Albert & Barabási, 2002). The structure of English thesaurus networks and other semantic networks was first studied by M. Steyvers and J. B. Tenenbaum (2005, 2001)[3]. Albert & Barabási (2002) reported results on the Merrian-Webster dictionary by Yook *et al*. Steyvers & Tenenbaum's study was carried out on the 1911 version of the Roget's thesaurus. In a latter paper, Motter *et al.* (2002) presented a study of an English thesaurus network based on the Moby Thesaurus[4]. This thesaurus is a slightly expanded version of the 1911 Roget's Thesaurus, so the properties of the Moby Thesaurus does not add anything substantially new to previous works. A reduced version of the full Roget' thesaurus was also analyzed in Newman (2003).

WorNet is a special kind of linguistic network. WordNet is an attempt from psycholinguistics theory to define word meaning and model not only word-meaning associations but also meaning-meaning associations. Thesaurus networks contain vaguely defined links. WordNet is an improved and extended thesaurus. It is improved, since it seeks to define word-meaning associations in a more precise way and, it is extended, since it includes various types of information that is not available in standard thesaurus. WordNet (Steyvers & Tenenbaum, 2005, 2001; Sigman & Cecchi; 2002) was developed by George Miller and colleagues (Miller *et al.*, 1990; Miller, 1995). The network is formed by two types of vertices: words and concepts. Words can be connected to each other through a variety of relationships such as synonymy and antonymy. Words

[3] It is worth mentioning that Steyver's and Tenenbaum's work has not been published untill very recently even though it was written long time ago. This work that has not been properly referenced in the literature. In Barabási and Albert (2002), the authors cite the first version of the paper, prior to Steyvers & Tenenbaum (2001) that later appeared as Steyvers and Tenenbaum (2001). As far as we know, Steyvers and Tenenbaum study remains the only study where more than one kind of linguistic network has been studied simultaneously.

[4] The Project Gutenberg Etext of Moby Thesaurus II by Grady Ward. Available at ftp://ibiblio.org/pub/docs/books/gutenberg/etext02/mthes.zip.

that are synonymous form a *synset* (from synonymy set). Each concept has an associated synset. Concepts are connected by relationships such as hypernymy (*maple* and *tree*) and meronymy (*bird* and *beak*). WordNet contains only four parts-of-speech: verbs, nouns, adverbs and adjectives. The study in Steyvers & Tenenbaum (2005, 2001) is based on all four word types and mixes words with concepts as vertices. In contrast, the study by Sigman & Cecchi (2002) is focused on nouns and vertices are only nominal concepts.

Word association networks are constructed from data from a specific psychological experiment. Participants have to respond quickly by freely generating an associate word (the response) to a cue word given as input (the stimulus). All the stimulus-response associations are recorded. The word association network is formed by all words in the experiment (stimuli and responses) and directed links are drawn from each stimulus to the corresponding responses, assuming that a link is oriented from the stimulus to the response (Steyvers & Tenenbaum, 2005, 2001; Capocci *et al.*, 2005). See Fig. 1.

Word co-occurrence networks (Ferrer i Cancho & Solé, 2001; Dorogovtsev & Mendes, 2001, 2003a; Milo *et al.*, 2004) are constructed from a corpus, i.e. a large collection of sentences. The idea behind them is that near words in a sentence tend to be syntactically related. The word co-occurrence network of a corpus contains all words in the corpus as vertices. A pair of words is linked if both words have appeared at least once at a distance smaller or equal than $D$ in the corpus. $D$ is called the window size. Here the distance between a word $x$ and word $y$ in a sentence (assuming that $x$ appeared first) is the number of words inbetween $x$ and $y$ plus 1. That is a Euclidean distance because it is a measure in a one-dimensional Euclidean space. Table 2 shows the Euclidean distance between words in the sentence

"She loved me for the dangers I had passed." (1)

About 90% of syntactic relationships take place at distance lower or equal than two (Ferrer i Cancho, 2004, to appear), but word co-occurrence networks lack a linguistically precise definition of link and fail in capturing the characteristic long-distance correlations of words in sentences (Chomsky, 1957). The amount of co-occurrences captured (per sentence) that are not syntactic dependencies is about 50% with a window $D$=2 and about 30% with a window $D$=1 (Ferrer i Cancho, *et al.* 2004)[5]. Therefore, the amount of syntactically incorrect links captured (per sentence) in Ferrer i Cancho & Solé (2001), where $D$=2, is about 70%. That amount is about 30% in Milo *et al.* (2004), where $D$=1. Word co-occurrence networks should be seen as rough approximations. Syntactic dependency networks overcome the problems above and will be the subject of

[5] The values where estimated using a Romanian syntactic dependency corpus. See Ferrer i Cancho (2004), Ferrer i Cancho *et al.*, (2004), the web link http://phobos.cs.unibuc.ro/roric/DGA/dga.html for further details.

the next two sections. For more details about the other linguistics networks, please follow the references.

## THE SYNTACTIC DEPENDENCY STRUCTURE OF SENTENCES

Dependency grammar (Melčuk, 1988; Melčuk, 2002) is a formalism defining the structure of a sentence by means of a graph which is generally a tree as in Fig. 2. Links between pairs of words are syntactic dependencies. Most of the links are directed and the arc usually goes from the head to the modifier word. In some cases, such as coordination, there is no clear direction (Melčuk, 2002). As in previous studies, we neglect link direction because it is simpler (Ferrer i Cancho *et al.* 2004) or irrelevant if one focuses on the length of sentence arcs (Ferrer i Cancho, 2004, 2005d, to appear).

Rather than the statistical properties of a sentence structure, which we assume here to be a tree (an acyclic connected graph; see Bollobás (1998) for an introduction to standard graph theory), we are interested in the relationship between the position of words in a sentence and the structure of the sentence. It is surprising that about 90% of the links of a sentence are formed between first or second neighbours within sentences. In fact, that locality phenomenon is not expected from a random arrangement of words in sentences (Ferrer i Cancho, 2004; to appear). It is well-known in cognitive science that the distance between syntactically related items within a sentence is a measure of the cost involved in handling that relationship by the brain (Grodner & Gibson, 2005; Gibson & Pearlmutter, 1998). Here cost means the amount of brain resources needed. We define $\langle d \rangle$ as the mean Euclidean distance between pairs of syntactically related words in a sentence. A further statistical analysis of syntactic dependency corpora strongly suggests that $\langle d \rangle$ is minimized or constrained to a small value (Ferrer i Cancho, 2004). Constraining $\langle d \rangle$ to a small value, predicts an exponential distribution of distances between syntactically related items of a sentence that is consistent with the real distribution (Ferrer i Cancho, 2004). The predictions do not stop here.

It is well known that syntactic dependency links do not generally cross when drawn over a sentence. That universal property of sentence structures is a fundamental ingredient of projectivity (Melčuk, 1988). Fig. 2 shows the syntactic dependency structure of Sentence 1, where no crossings are found. Fig. 3 shows that many crossings appear when scrambling the words in sentence 1. Why do syntactic links generally not cross? Minimizing $\langle d \rangle$ can explain the absence, in general, of crossings in sentence structures (Ferrer i Cancho, 2005d). The cost of distant links produces a tension that precludes crossings. All together, the limited resources of the brain translate into heavy constraints on the length of links, which emerge as an exponential distribution of distances and the exceptional nature of link crossings in sentences.

## THE GLOBAL SYNTACTIC DEPENDENCY STRUCTURE

Syntactic dependency networks (Ferrer i Cancho *et al*., 2004, 2005b) are constructed from a corpus as word co-occurrence networks. Here the structure of every sentence is specified using the dependency grammar formalism. The syntactic dependency network of a corpus contains all words in that corpus. A pair of words is linked if the words involved have appeared syntactically linked at least once in a sentence of the corpus. Thus, a global syntactic dependency network is constructed by cumulating sentence structures from a corpus. That global network is an emergent property of sentence structures. In fact, the statistical properties of the global syntactic dependency network cannot be explained by the statistical properties of the structure of sentences (Ferrer i Cancho *et al.*, 2004). Fig. 4 shows a global network obtained from a Romanian dependency corpus. The drawing is messy and obtaining a nice layout not only depends on the visualization technique but also on the structure of the network. For certain networks obtaining a nice layout seems to be an inherently difficult problem (for instance, when there are too many links with regard to the number of vertices). What is important here is that we cannot always rely on visualization techniques for gaining knowledge about the structure of a network and visual inspection is limited (in the sense that drawings are often messy and the global properties are not always easy to read from the plot). In contrast, the statistical analysis techniques developed by physicists (Albert & Barabási, 2002; Newman, 2003) can provide a precise characterization and unravel fundamental properties that are hidden to the eye.

In the global network, the Euclidean distance between words in sentences is lost. We can focus our attention on the network distance, a distance that is defined in the network space. We are interested in the minimum distance between two words $u$ and $v$ in the network space. That distance is defined as the minimum amount of links that need to be crossed in order to reach $v$ starting from $u$. Table 3 shows how to calculate minimum network distances in a real case. For the sake of clarity, the case is a sentence structure (we could have used the global network but the plot in Fig. 4 is messy). The analysis of network distances in syntactic dependency networks shows the presence of the small-world phenomenon: despite of the large amount of vertices in the network, the distance between them is surprisingly small. For the Romanian network in Fig. 4 (Ferrer i Cancho *et al.*, 2004, 2005b), with 5,563 words, the minimum number of edges needed for reaching any word of the network is 3.4 on average. Starting from any word in the network, the remaining words are reached in about three steps (on average) which is a very small quantity compared to the total number of words. Here, the reasons for surprise are moderate with regard to the small Euclidean distance between syntactically linked words, since a global network formed by the same amount of elements and connections but forming links by choosing pairs of words at random would give a similar small network distance

(Watts & Strogatz, 1998). Another essential property of global syntactic dependency networks is a heterogeneous degree distribution. The degree is the number of connections of a vertex (e.g. a word). Roughly speaking, many words have a few connections but the proportion of words with many links is significant. The degree distribution takes the mathematical form of a power law, which is very different from an exponential distribution, where the probability that a word has many connections could be neglected. See Newman (2005) for an accessible explanation of what a power law is and examples of systems following and not following that kind of distribution. Interestingly, the small-world phenomenon is found in all the linguistic networks seen so far and a power degree distribution is found in most of them. Due to space limits, we cannot review all the statistical properties found in syntactic dependency networks and the other linguistic networks discussed.

The power degree distribution that is found could be explained by many models (Bornholdt & Ebel, 2001; Dorogovtsev & Mendes, 2003b; Newman, 2003). Given the small network distance between words in the global network, it is tempting to think that the degree distribution could be the outcome of a network distance minimization process producing a that kind of distribution as the model in Ferrer i Cancho & Solé (2003), as Euclidean distance minimization seems to work at the sentence level. We will show that this is not necessary.

Although a power degree distribution of that kind could be generated by many other models, Occam razor's favours a particular track. The relationship between word frequency and word degree is approximately linear (Ferrer i Cancho *et al.*, 2004), so the distribution of word degrees could be a consequence of the distribution of words frequencies. Interestingly, the word frequency distribution, known as Zipf's law for word frequencies (Zipf, 1935, 1949), is a power law with approximately the same exponent as the word degree distribution. Following that explanatory track, recent models explain the word degree distribution as a side-effect of, roughly speaking, the associations of words with meanings in a communication system (Ferrer i Cancho *et al.*, 2005a; Ferrer i Cancho, 2005c). If word degree is a consequence of Zipf's law for word frequencies, a pressing question is: what the origin of that law is?

Recent models strongly suggest that Zipf's law could be the outcome of very general communication principles: roughly speaking, maximizing the information transfer from words to meanings while the cost of word use is minimized (Ferrer i Cancho, 2005b,2005c; Ferrer i Cancho & Solé, 2003) or constraining the ambiguity of words (Ferrer i Cancho, 2005a). Of special interest here is that the cost of word use is imposed by the negative correlation between the frequency of a word and its availability, the so-called word frequency effect (Akmajian *et al.*, 1995).

## DISCUSSION

Global syntactic networks and the ordering of words in sentences above seem to be shaped by brain limitations. Regarding global syntactic networks, the limited capacity of the brain translates into heavy constraints on the frequency of words, which emerge as Zipf's law for word frequencies and, presumably, as a power degree distribution. Interestingly, three different languages (Romanian, Czech and German) were studied and common traits where founds despite of the limitations of the data. The work by Ferrer i Cancho *et al.* (2004) is, as far as we know, the only linguistic network study where more than one language has been considered. The statistical regularities found suggest they could be universal. Besides, it is the only linguistic network study where a language different than English is considered. That ethnocentric bias should be overcome in the future.

We have seen that brain constraints may shape the ordering of words within sentences and the structure of global syntactic dependency networks. The statistical properties of global syntactic dependency networks have many implications. The small-world phenomenon can explain why mental navigation through the web of words is easy: one can start in any word of the network and reach the remaining words in a few steps. The heterogeneous degree distribution may explain why the capacity to produce complex sentences is severely affected in agrammatism, a kind of aphasia (Caplan, 1997). Agrammatism is characterized by the omission of function words. The most connected vertices in a network are called hubs (Albert & Barabási, 2002; Newman, 2003). The hubs of global syntactic dependency networks are function words. In general, networks with a power degree distribution are very robust against the disconnection of the low degree vertices but very sensitive to the disconnection of hubs (Jeong *et al.*, 2002). When hubs are removed, the network breaks into pieces (Albert *et al.*, 2000; Albert *et al.*, 2001). Interestingly, the structure of global syntactic dependency networks mirrors the structure of the brain. It is obvious that the brain is made of millions of neurons connected through synapses but the similarities go beyond mere physical resemblance. The activation of different brain areas shows the small-world phenomenon and a power degree distribution (Eguíluz *et al*. 2005; Grinstein & Linsker, 2005). The strong coincidence questions the suitability of classic phase structure models (Chomsky, 1957) and later developments (Chomksy, 1995; Uriagereka, 1998) for modelling human language and suggest that a natural approach to the structure of language would be closer to syntactic dependency based formalisms. While no one has ever find a rewriting rule in the brain of a human, the web organization of the brain at many levels, with linguistic networks on top, cannot be denied.

The statistical tools developed by physicists for studying networks could be of great help in the quest for new linguistic universals. Some candidates for global syntactic dependency networks are the small-world phenomenon and a power-

degree distribution. If it turned out that the degree of a word is a consequence of its frequency, as hypothesized above, the case that the second candidate was an actual universal would not be surprising since Zipf's law for word frequencies is an apparently universal property of world-languages (Naranan & Balasubrahmanyan, 1996) and recent models (Ferrer i Cancho, 2005b, Ferrer i Cancho & Solé, 2003) suggest it should be so even for the languages where the presence of Zipf's law has not yet been checked.

In sum, the recent developments of network theory offer new possibilities for defining the universal properties of world languages and understanding their origin and implications.

## ACKNOWLEDGMENTS

We thank Brita Elvevåg for helping us to improve the English. We are grateful to Sergi Valverde for the network drawing of Fig. 4 and Francesco Rao for technical assistance. This work was supported by the ECAgents project, funded by the Future and Emerging Technologies program (IST-FET) of the European Commission under the EU RD contract IST-1940. The information provided is the sole responsibility of the author and does not reflect the Commission's opinion. The Commission is not responsible for any use that may be made of the data appearing in this publication.

## REFERENCES

Akmajian, A., Harnish, R. M., Demers, R. A. & Farmer, A. K. (1995). Linguistics. An introduction to language and communication. Cambridge, MA: MIT Press.

Albert, R. Jeong, H. & Barabási, A.-L. (2000). Error and attack tolerance of complex networks. Nature 406, 378-382.

Albert, R. Jeong, H. & Barabási, A.-L. (2001). Correction: Error and attack tolerance of complex networks. Nature 409, 542.

Albert, R. & Barabási, A.-L. (2002). Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47-97.

Balasubrahmanyan, V. K. & Naranan, N. (1996). Quantitative linguistics and complex systems studies. Journal of Quantitative Linguistics 3, 177-228.

Bollobás, B. (1998). Modern graph theory. New York: Springer-Verlag.

Bornholdt, S. & Ebel, H. (2001). World Wide Web scaling from Simon's 1955 model. Physical Review E 64, 035104 (R).

Capocci, A., Servedio, V.D.P, Caldarelli, G. & Colaiori, F. (2005). Detecting communities in large networks. Physica A 352, 669-676.

Caplan, D. (1987). Neurolinguistics and linguistic aphasiology. New York: Cambridge University Press.

Chomsky, (1957). Syntactic structures. New York: Mouton.

Chomsky, (1995). The minimalist program. Cambridge, MA: MIT Press.

Dorogovtsev, S.N. & Mendes, J.F.F. (2001). Language as an evolving word web. Proceedings of the Royal Society of London B 268, 2603-2606.

Dorogovtsev, S.N. & Mendes, J.F.F. (2003a). Accelerated growth of networks. In: Handbook of graphs and networks. From the genome to the Internet, Bornholdt, S. and Schuster, G. H. (eds.). Winheim, Wiley-VCH, pp. 318-339.

Dorovotsev, S. N. & Mendes, J. F. F. (2003b). Evolution of networks – From biological nets to the Internet and the WWW. Oxford, Oxford University Press.

Dorogovtsev, S.N. & Mendes, J.F.F. (2002). Evolution of networks. Advances in Physics 51, 1079-1187.

Eguíluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M. & Apkarian, A. V. (2005). Scale-free brain functional networks. Physical Review Letters 94, 018102.

Ferrer i Cancho, R. & Solé, R. V. (2001). The small world of human language. Proceedings of the Royal Society of London B 268, 2261-2265.

Ferrer i Cancho, R. & Solé, R. V. (2003). Optimization in complex networks. In: Statistical Mechanics of complex networks, Pastor-Satorras, R. *et al*. (eds.). Lecture Notes in Physics 625, 114-125. Berlin: Springer.

Ferrer i Cancho, R. & Solé, R. V. (2003). Least effort and the origins of scaling in human language. Proceedings of the National Academy of Science USA 100, 788-791.

Ferrer i Cancho, R., Solé, R.V. and Köhler, R. (2004). Patterns in syntactic dependency networks. Physical Review E 69, 051915.

Ferrer i Cancho, R. (2004). The Euclidean distance between syntactically linked words. Physical Review E 70, 056135.

Ferrer i Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. Physica A 345, 275-284.

Ferrer i Cancho, R. (2005b). Zipf's law from a communicative phase transition. European Physical Journal B 47, 449-457.

Ferrer i Cancho, R. (2005c). When language breaks into pieces. A conflict between communication through isolated signals and language. Biosystems (in press).

Ferrer i Cancho, R. (2005d). Why do syntactic links not cross? Submitted.

Ferrer i Cancho, R. (to appear). The locality of syntactic dependencies.

Ferrer i Cancho, R., Riordan, O. & Bollobás, B (2005a). The consequences of Zipf's law for syntax and symbolic reference. Proceedings of the Royal Society of London B 272, 561-565.

Ferrer i Cancho, R., Capocci, A. & Caldarelli, G. (2005b). Spectral methods cluster words of the same part-of-speech in a syntactic dependency network. cont-mat/0504165.

Gibson, E., & Pearlmutter, N. J. (1998). Constraints on sentence comprehension. Trends in Cognitive Sciences, 2, 262-268.

Grinstein, G. & Linsker, R. (2005). Synchronous neural activity in scale-free network models versus random network models. Proceedings of the National Academy of Sciences USA 102, 9948-9953.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. Cognitive Science, 29, 261-290.

Jeong, H., Mason, S.P., Barabási, A.-L. & Oltvai, Z. N. (2002). Lethality and centrality in protein networks. Nature 411, 41-42.

Kinouchi, O., Martinez, A. S., Lima, G.F., Lourenço, G. M. & Risau-Gusman, S. (2002). Deterministic walks in random networks: an application to thesaurus graphs. Physica A 315, 665-676.

Melčuk, I. (1998). Dependency Syntax. New York: SUNY.

Melčuk, I. (2002). Language: dependency. In. N. J. Smelser and P. B. Baltes, (eds.). International Encyclopedia of the Social and Behavioral Sciences, pp 8336-8344. Oxford: Pergamon.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. (2004). Superfamilies of evolved and designed networks. Science 303, 1538-1542.

Miller, G. A., Beckwith, R. Fellbaum, C., Gross, D. & Miller, K.J. (1990). WordNet: an on-line lexical database. International Journal of Lexicography 3, 235-244.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM 38, 39-41.

Motter, A. E., de Moura, A.P.S., Lai, Y.-C. & Dasgupta, P. (2002). Topology of the conceptual network of language. Physical Review E 65, 065102(R).

Newman, M. E. J. (2003). The structure and function of complex networks. SIAM Review 45, 167-256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323-351.

Sigman, M. & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon 99, 1742-1747.

Steyvers, M. & Tenenbaum, J. B. (2001). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. cond-mat/0110012.

Steyvers, M. & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognitive Science 29, 41-78.

Uriagereka, J. (1998). Rhyme and Reason. An introduction to minimalist syntax. Cambridge, MA: MIT Press.

Watts, D. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature 393, 440-442.

Zipf, G. K. (1935). The psycho-biology of language: an introduction to dynamic philology. Boston, MA: Houghton-Mifflin.
Zipf, G. K. (1949). Human behavior and the principle of least effort. Reading: Addison-Wesley.

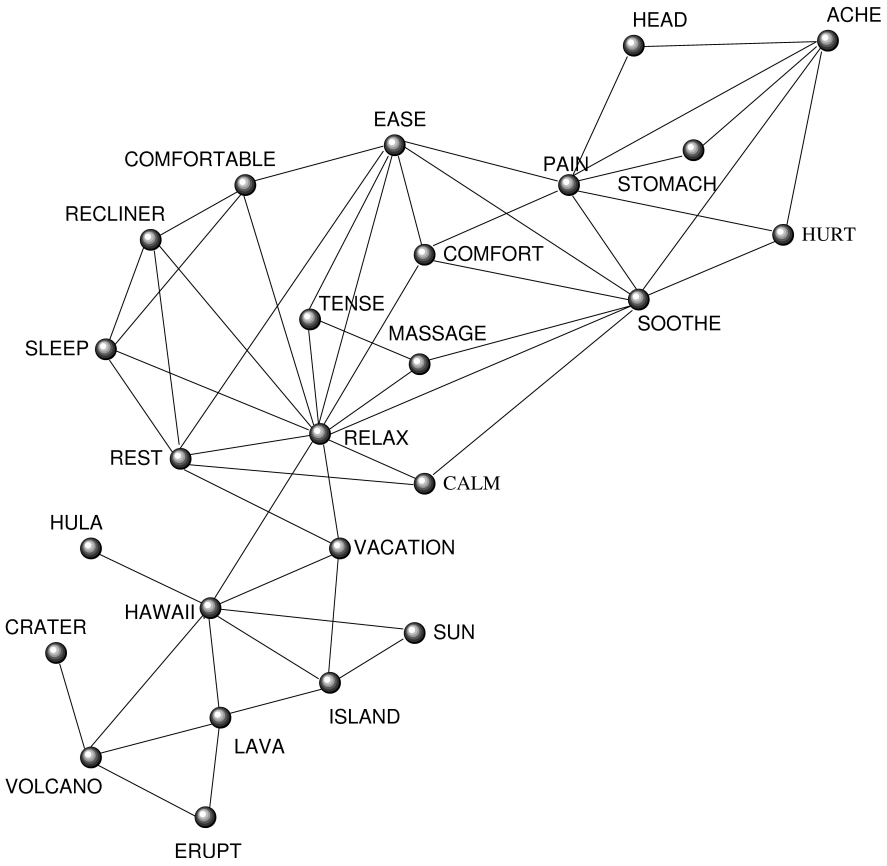


Fig. 1. A subgraph of the word association network in Steyvers and Tenenbaum (2005). Vertices are words and two words are linked if one of them has been given as response to the other (link direction is neglected here). The figure has been redrawn from Steyvers & Tenenbaum (2005, 2001).

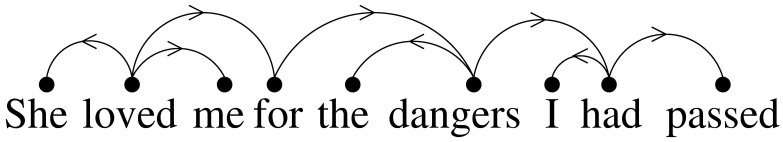


Fig 2. The syntactic structure of the sentence *'She loved me for the dangers I had passed'* following the conventions in (Melčuk, 1989). Here vertices are words and arcs stand for syntactic dependencies. Following those conventions, arcs go from heads to modifiers. The pronoun *'she'* and the verb *'loved'* 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.

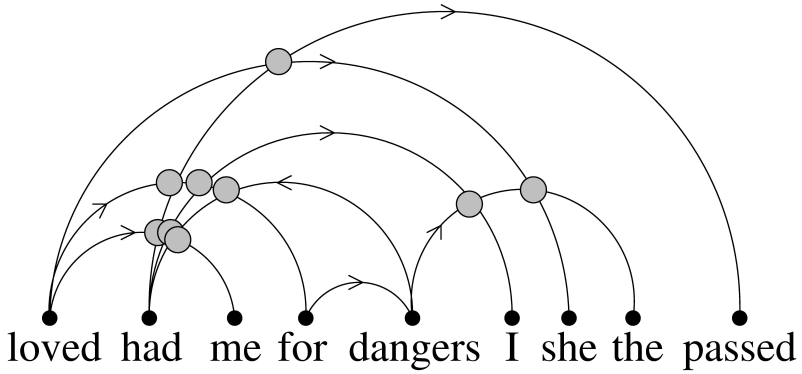


Fig. 3. The structure of the sentence in Fig. 1 after having scrambled the words. Gray circles indicate edge crossings.

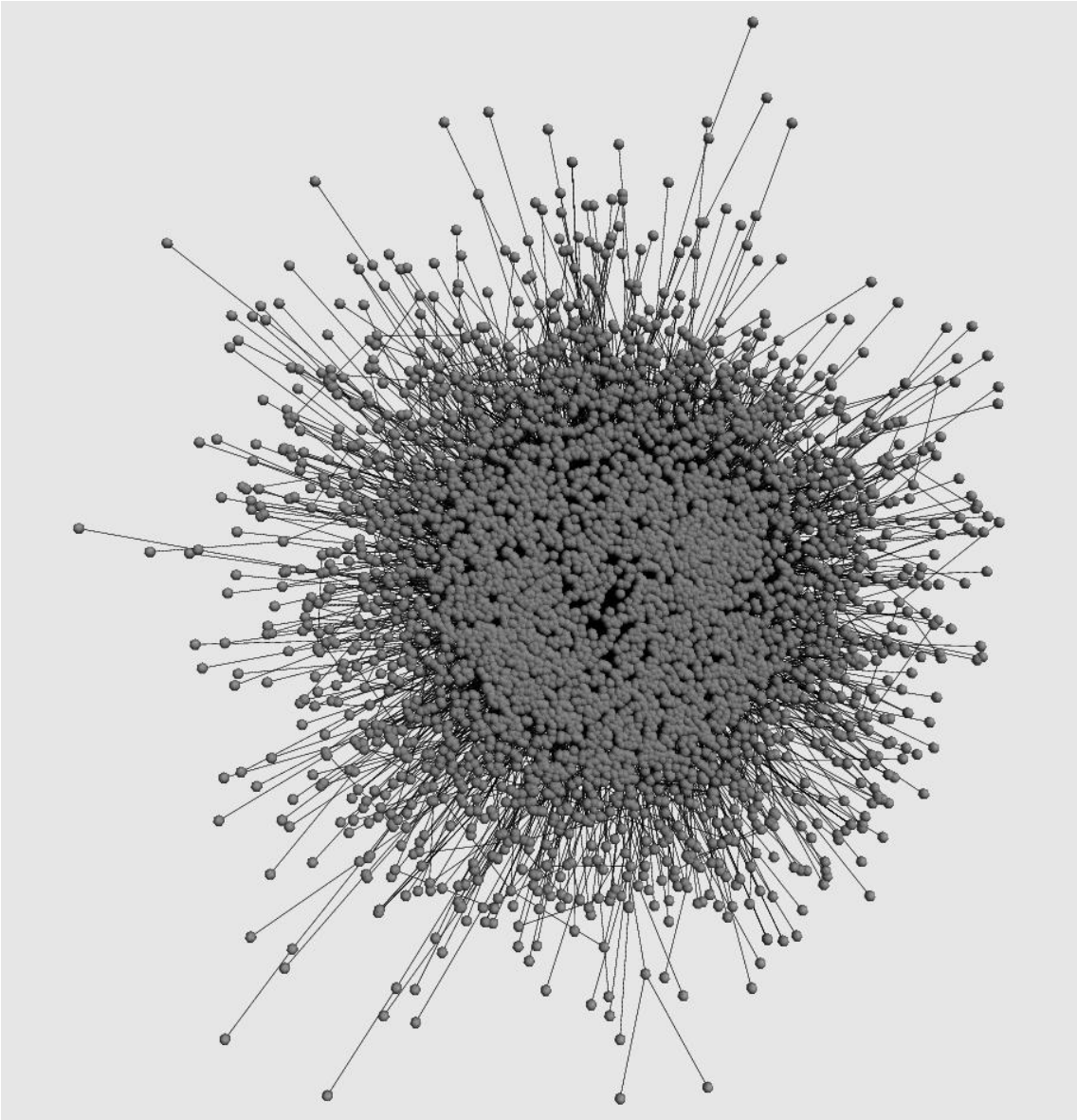

Fig. 4. The global syntactic dependency network of a Romanian corpus. (drawing by S. Valverde). Vertices are Romanian words and connections indicate syntactic dependencies. The network has 5,563 vertices.

| Root | Related words |
|---|---|
| *acclamation* | *acclaim, accord, accordance, agreement, agreement of all, applause, big hand, burst of applause, cheer, horus, clap, clapping, clapping of hands, common assent, common consent, concert, concord, concordance, concurrence, consensus, consensus gentium, consensus of opinion, consensus omnium, consent, consentaneity, eclat, encore, general acclamation, general agreement, general consent, general voice, hand, handclap, handclapping, harmony, like-mindedness, meeting of minds, mutual understanding, one accord, one voice, ovation, plaudit, plaudits, popularity, round of applause, same mind, single voice, thunder of applause, total agreement, unanimity, unanimousness, understanding, unison, universal agreement.* |
| *acclimate* | *acclimatize, accommodate, accustom, adapt, adjust, break, break in, case harden, condition, confirm, domesticate, domesticize, establish, familiarize, fix, gentle, habituate, harden, housebreak, inure, naturalize, orient, orientate, season, tame, toughen, train, wont.* |

Table 1. Two consecutive entries in the Moby thesaurus: *acclamation* and *acclimate*. The Moby thesaurus is based on the Roget's thesaurus 1911 edition.

| $d(u,v)$ | | $v$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *she* | *loved* | *me* | *for* | *the* | *dangers* | *I* | *had* | *passed* |
| $u$ | $\pi(v)$ / $\pi(u)$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| *she* | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| *loved* | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| *me* | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| *for* | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| *the* | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 | 4 |
| *dangers* | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
| *I* | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 | 2 |
| *had* | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 | 1 |
| *passed* | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 2 |

Table 2. Matrix of Euclidean distance (in words) between linked words in the sentence "*She loved me for the dangers I had passed*". Link direction is neglected. The distance between syntactically linked words appears with a gray background. The matrix is symmetric because $d(u,v)=d(v,u)$.

| $\delta(u,v)$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $v$ / $u$ | *she* | *loved* | *me* | *for* | *the* | *dangers* | *I* | *had* | *passed* |
| *she* | 0 | 1 | 2 | 2 | 4 | 3 | 5 | 4 | 5 |
| *loved* | 1 | 0 | 1 | 1 | 3 | 2 | 4 | 3 | 4 |
| *me* | 2 | 1 | 0 | 2 | 4 | 3 | 5 | 4 | 5 |
| *for* | 2 | 1 | 2 | 0 | 2 | 1 | 3 | 2 | 3 |
| *the* | 4 | 3 | 4 | 2 | 0 | 1 | 3 | 2 | 3 |
| *dangers* | 3 | 2 | 3 | 1 | 1 | 0 | 2 | 1 | 2 |
| *I* | 5 | 4 | 5 | 3 | 3 | 2 | 0 | 1 | 2 |
| *had* | 4 | 3 | 4 | 2 | 2 | 1 | 1 | 0 | 1 |
| *passed* | 5 | 4 | 5 | 3 | 3 | 2 | 2 | 1 | 0 |

Table 3. Matrix of network minimum distance (in edges) between pairs of words in the sentence "*She loved me for the dangers I had passed*". $\delta(u,v)$ is the minimum network distance between words $u$ and $v$. $\delta(u,v)$ is the minimum amount of links that need to be crossed for reaching $v$ starting from $v$. Link direction is neglected and hence the matrix is symmetric ($\delta(u,v)=\delta(v,u)$).