10.12871/CLICIT2014156

The Importance of Being *sum*. Network Analysis of a Latin Dependency Treebank

Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
marco.passarotti@unicatt.it

Abstract

English. Network theory provides a suitable framework to model the structure of language as a complex system. Based on a network built from a Latin dependency treebank, this paper applies methods for network analysis to show the key role of the verb *sum* (*to be*) in the overall structure of the network.

Italiano. *La teoria dei grafi fornisce un valido supporto alla modellizzazione strutturale del sistema linguistico. Basandosi su un network costruito a partire da una treebank a dipendenze del latino, l'articolo applica diversi metodi di analisi dei grafi, mostrando l'importanza del ruolo rivestito dal verbo* sum *(*essere*) nella struttura complessiva del network.*

1 Introduction

Considering language as a complex system with deep relations between its components is a widespread approach in contemporary linguistics (Briscoe, 1998; Lamb, 1998; Steels, 2000; Hudson, 2007). Such a view implies that language features complex network structures at all its levels of analysis (phonetic, morphological, lexical, syntactic, semantic).

Network theory provides a suitable framework to model the structure of linguistic systems from such a perspective. Network theory is the study of elements, called *vertices* or *nodes*, and their connections, called *edges* or *links*. A complex network is a (un)directed graph $G(V, E)$ which is given by a set of vertices V and a set of edges E (Ferrer i Cancho, 2010).

Vertices and edges can represent different things in networks. In a language network, the vertices can be different linguistic units (for instance, words), while the edges can represent different kinds of relations holding between these units (for instance, syntactic relations).

So far, all the network-based studies in linguistics have concerned modern and living languages (Mehler, 2008a). However, times are mature enough for extending such approach also to the study of ancient languages. Indeed, the last years have seen a large growth of language resources for ancient languages. Among these resources are syntactically annotated corpora (treebanks), which provide essential information for building syntactic language networks.

2 From a Dependency Treebank to a Syntactic Dependency Network

For the purpose of the present study, we use the *Index Thomisticus* Treebank, a Medieval Latin dependency treebank based on the works of Thomas Aquinas (IT-TB; http://itreebank.marginalia.it; Passarotti, 2011). Presently, the IT-TB includes around 200,000 nodes in approximately 11,000 sentences. For homogeneity reasons, in this work we consider the subset of the IT-TB that features the in-line

annotation of the text of the *Summa contra Gentiles* (entire first book and chapters 1-65 of the second one) for a total of 110,224 nodes.

Automatic data cleaning was performed before building the network, by excluding punctuation marks, function words and elliptical dependency relations from the input data. Then, the method developed by Ferrer i Cancho et alii (2004) was applied to build the network.

According to this method, a dependency relation appearing in the treebank is converted into an edge in the network. The vertices of the network are lemmas. Two lemmas are linked if they appear at least once in a modifier-head relation (dependency) in the treebank.

Then a syntactic dependency network is constructed by accumulating sentence structures from the treebank. The treebank is parsed sentence by sentence and new vertices are added to the network. When a vertex is already present in the network, more links are added to it.

The result is a syntactic dependency network containing all lemmas and all dependency relations of the treebank. All connections between particular lemmas are counted, which means that the graph reflects the frequency of connections. The network is an emergent property of sentence structures (Ferrer i Cancho, 2005; Ferrer i Cancho et al., 2004), while the structure of a single sentence is a subgraph of the global network (Bollobás, 1998).

The free software Cytoscape was used for network creation and computing (Shannon et al., 2003; Saito et al., 2012).

Figure 2 presents the syntactic dependency network of the subset of the IT-TB used in this work. Vertices and edges are arranged according to the Edge-weighted Spring Embedded layout setting provided by Cytoscape (Kohl et al., 2011). Edges are weighted by frequency, the most central relations in the network being those most frequent in the treebank.

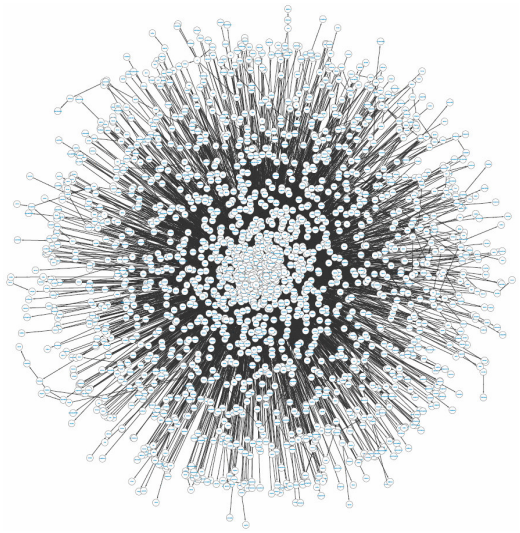

Figure 1. The network of the IT-TB

The drawing in figure 1 is messy and not much informative. In order to both analyze and categorize the network, we use a number of topological indices that are able to unravel fundamental properties of the network that are hidden to the eye.

3 Topological Indices

Most complex networks are characterized by highly heterogeneous distributions (Newman, 2005a). This property means that there are many vertices having a few connections and a few vertices with a disproportionately large number of connections. The most connected vertices in a network are called *hubs* (Albert & Barabási, 2002; Newman, 2003).

In network analysis, the centrality of a vertex is a topological index that measures its relative importance within a graph. We use two measures of centrality ('betweenness' and 'closeness') to calculate the importance of a vertex in a syntactic dependency network, i.e. to find hubs in the network. The higher are betweenness and closeness centralities of a vertex, the more important the vertex is in the network.

The *betweenness centrality* of a vertex *v, g(v)*, is a measure of the number of minimum distance (or "shortest") paths running through *v* (Ferrer i Cancho et al., 2004).

Closeness centrality. In a network, the length of the shortest paths between all pairs of vertices is a natural distance metric. The "farness" of a vertex *s* is the sum of its distances to all other vertices, and its "closeness" is the inverse of the farness (Sabidussi, 1966). Thus, the more central a vertex is, the lower is its total distance to all other vertices. Closeness centrality is a measure of how long it takes to spread information from *s* to all other vertices sequentially in the network (Newman, 2005b; Wuchty & Stadler, 2003).

Further, we use the following topological indices in order to categorize a syntactic dependency network by evaluating its complexity (Mehler, 2008b).

The so-called *degree* of a vertex *s* is the number of different relations holding between *s* and other vertices in the network. The *average degree* $\varepsilon(G)$=*edges/vertices* of a graph *G* is the proportion of edges with respect to the number of vertices.

Clustering coefficient is the probability that two vertices that are neighbours of a given vertex are neighbours of each other (Solé et al., 2010). In other words, it is a measure of the relative frequency of triangles in a network.

Average path length. Path length is defined as the average minimal distance between any pair of vertices (Solé et al., 2010). The average path length *d* is defined as the average shortest distance between any pair of vertices in a network.

Together with the clustering coefficient, the average path length of a graph *G* constitutes the 'small-world model' of Watts & Strogatz (1998), which has proved to be an appropriate model for many types of networks (like, for instance, biological and social ones). If a network has a high clustering coefficient but also a very short average path length in comparison to random graphs with the same number of vertices, it is a small-world network.

4 Hubs in the IT-TB Network

For each vertex in the IT-TB network, we calculated its betweenness and closeness centralities using the Cytoscape app CytoNCA (http://apps.cytoscape.org/apps/cytonca).

Table 1 presents the rates of the centrality measures of the first five lemmas in the IT-TB network ranked by betweenness centrality. The table reports also the degree for each lemma.

R.	Lemma	Betw. C.	Clos. C.	Deg.
1	*sum* (*to be*)	1793719.9	0.2822	1095
2	*dico* (*to say*)	324728.16	0.2558	401
3	*possum* (*can*)	307137.8	0.2581	464
4	*habeo* (*to have*)	214495.38	0.2535	351
5	*facio* (*to make*)	146891.89	0.2507	289

Table 1. Results on centrality measures

Although some lemmas are differently ranked according to different centrality measures (for instance, *dico* is second by betweenness centrality, but it is third by both closeness centrality and degree), *sum* remains always first. This shows that *sum* is the "most hub" among the hubs of the IT-TB network.

Hubs are the key components of the complexity of a network. They support high efficiency of network traversal, but, just because of their important role in the web, their loss heavily impact the performance of the whole system (Jeong et al., 2002). If the most highly connected vertices are removed, the network properties change radically and the network breaks into fragments, sometimes even promoting a system's collapse (Albert & Barabási, 2000).

Following its status of most hub vertex in the IT-TB network, we removed the vertex of *sum* and of all its direct neighbours from the network. Further, we removed all those vertices that become isolated in the network after such a removal is applied (i.e. those with degree = 0; in total: 702 vertices). Figure 2 presents the subnetwork that results from these modifications.

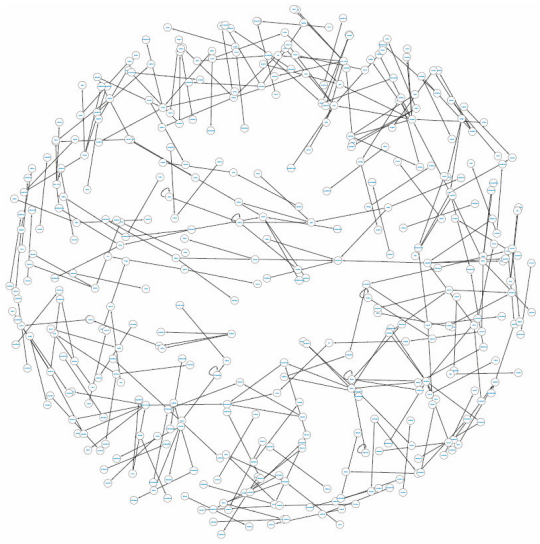

Figure 2. The IT-TB no-*sum* subnetwork

The counterpart of the subnetwork in figure 2 is the subnetwork formed only by the vertex of *sum* and its direct neighbours (figure 3).

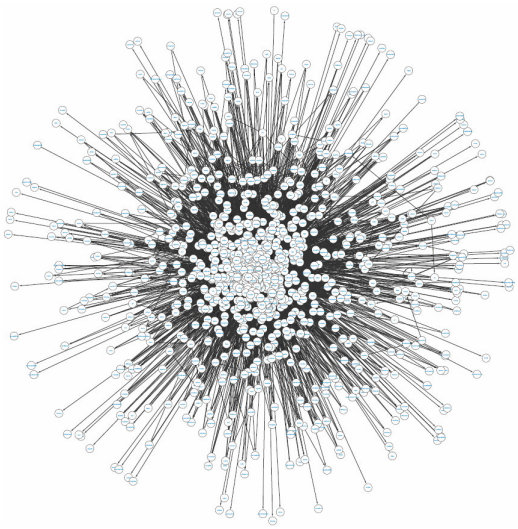

Figure 3. The IT-TB *sum*-only subnetwork

While figure 2 shows that removing the vertex of *sum* and those of its direct neighbours makes the network lose its connecting core, figure 3 presents a very much connected subnetwork.

In order to evaluate the role of *sum* in the network beyond the graphical layout of the subnetworks, we calculated the above mentioned topological indices of the full network of the IT-TB (1) and of the subnetworks reported respectively in figures 2 (2) and 3 (3). Table 2 presents the results.

	1	**2**	**3**
N. of vertices	2,198	398	1,098
N. of edges	19,031	301	15,486
Average degree	8.6583	0.7562	14.1038
Average path length	3.108	1.4883	2.5242
Clustering coefficient	0.247	0.081	0.352

Table 2. Results on topological indices

From the rates reported in table 2 it turns out that the subnetwork 2 is less small-world than 1 and 3, i.e. 2 is less connected and more fragmented than 1 and 3. This is shown by the clustering coefficient, which is dramatically lower in 2 than in 1 and 3. Although the average path length of 2 is shorter than 1 and 3, this is motivated by the much lower number of vertices in 2 than in 1 and 3, and not by the more small-worldness of 2. This is more clear if we look at the relation between the number of edges and the number of vertices in the networks. While in 1 and 3, the edges are much more than the vertices, in 2 the opposite holds, thus leading to much different average degrees.

The subnetwork 3 is even more small-world than 1. 3 is smaller than 1, as it results from removing a number of vertices from 1. This is why the average path length of 3 is shorter than 1. However, both the average degree and the clustering coefficient of 3 are higher than 1. It is worth noting that 3 includes, alone, half of the total of the vertices occurring in 1 and around 75% of the edges of 1: this shows that the vertex of *sum* is directly connected to half the vertices of the network and these connections cover most of those that occur in the IT-TB network.

5 Conclusion

While the most widespread tools for querying and analyzing treebanks give results in terms of lists of words or sequences of trees, network analysis permits a synoptic view of all the relations that hold between the words in a treebank. This makes network analysis a powerful method to fully exploit the structural information provided by a treebank, for a better understanding of the properties of language as a complex system with interconnnected elements.

References

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

R. Albert and A.L. Barabási. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74: 47-97.

B. Bollobás. 1998. *Modern Graph Theory*. Vol. 184 of *Graduate Texts in Mathematics*. Springer, New York.

T. Briscoe. 1998. Language as a Complex Adaptive System: Coevolution of Language and of the Language Acquisition Device. P. Coppen, H. van Halteren and L. Teunissen (eds.), *Proceedings of Eighth Computational Linguistics in the Netherlands Conference*. Rodopi, Amsterdam, 3-40.

R. Ferrer i Cancho, R.V. Solé and R. Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review* E69, 051915(8).

R. Ferrer i Cancho. 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. V. Levickij and G. Altmman (eds.), *Problems of quantitative linguistics*, 60-75.

R. Ferrer i Cancho. 2010. Network theory. P.C. Hogan (ed.), *The Cambridge Encyclopedia of the Language Sciences*. Cambridge University Press, Cambridge, 555-557.

R. Hudson. 2007. *Language Networks. The New Word Grammar*. Oxford University Press, Oxford.

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

M. Kohl, S. Wiese and B. Warscheid. 2011. Cytoscape: software for visualization and analysis of biological networks. *Methods in Molecular Biology*, 696: 291-303.

S.M. Lamb. 1998. *Pathways of the Brain. The Neurocognitive Basis of Language*. John Benjamins, Amsterdam.

A. Mehler. 2008a. Large text networks as an object of corpus linguistic studies. A. Lüdeling and K. Merja (eds.), *Corpus Linguistics. An International Handbook of the Science of Language and Society*. De Gruyter, Berlin - New York, 328-382.

A. Mehler. 2008b. Structural similarities of complex networks: A computational model by example of Wiki graphs. *Applied Artificial Intelligence*, 22: 619-683.

M.E.J. Newman. 2003. The Structure and Function of Complex Networks. *SIAM Review*, 45.2: 167-256.

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

M.E.J. Newman. 2005b. A measure of betweenness centrality based on random walks. *Social Networks*, 27: 39-54.

M. Passarotti. 2011. Language Resources. The State of the Art of Latin and the *Index Thomisticus* Treebank Project. M.S. Ortola (ed.), *Corpus anciens et Bases de données. «ALIENTO. Échanges sapientiels en Méditerranée», N°2*. Presses universitaires de Nancy, Nancy, 301-320.

G. Sabidussi. 1966. The centrality index of a graph. *Psychometrika*, 31: 581-603.

R. Saito, M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, S. Lotia, A.R. Pico, G.D. Bader and T. Ideker. 2012. A travel guide to Cytoscape plugins. *Naure Publishing Group*, 9(11): 1069-76.

P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11): 2498-504.

R.V. Solé, B. Corominas-Murtra, S. Valverde and L. Steels. 2010. Language networks: Their structure, function, and evolution. *Complexity*, 15(6): 20-26.

L. Steels. 2000. Language as a Complex Adaptive System. M. Schoenauer (ed.), *Proceedings of PPSN VI, Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 17-26.

D.J. Watts and S.H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393: 440-442.

S. Wuchty and P.F. Stadler. 2003. Centers of complex networks. *Journal of Theoretical Biology*, 223(1): 45-53.