ISSN 1122 - 1917

L'ANALISI LINGUISTICA E LETTERARIA

FACOLTÀ DI SCIENZE LINGUISTICHE E LETTERATURE STRANIERE
UNIVERSITÀ CATTOLICA DEL SACRO CUORE

1

ANNO XXIV 2016

EDUCATT - UNIVERSITÀ CATTOLICA DEL SACRO CUORE

L'ANALISI LINGUISTICA E LETTERARIA

FACOLTÀ DI SCIENZE LINGUISTICHE
E LETTERATURE STRANIERE

UNIVERSITÀ CATTOLICA DEL SACRO CUORE

1

ANNO XXIV 2016

PUBBLICAZIONE SEMESTRALE

L'ANALISI LINGUISTICA E LETTERARIA
Facoltà di Scienze Linguistiche e Letterature straniere
Università Cattolica del Sacro Cuore
Anno XXIV - 1/2016
ISSN 1122-1917
ISBN 978-88-9335-058-7

I contributi di questa pubblicazione sono stati sottoposti
alla valutazione di due Peer Reviewers *in forma rigorosamente anonima*

EDUCatt - Ente per il Diritto allo Studio universitario dell'Università Cattolica
Largo Gemelli 1, 20123 Milano | tel. 02.7234.2235 | fax 02.80.53.215
e-mail: editoriale.dsu@educatt.it (*produzione*); librario.dsu@educatt.it (*distribuzione*)
web: www.educatt.it/libri

Redazione della Rivista: redazione.all@unicatt.it | *web*: www.analisilinguisticaeletteraria.eu

Questo volume è stato stampato nel mese di luglio 2016
presso la Litografia Solari - Peschiera Borromeo (Milano)

Indice

Analisi d'opera

Recensioni e Rassegne

How Far Is Stanford from Prague (and vice versa)? Comparing Two Dependency-based Annotation Schemes by Network Analysis

Marco Passarotti

The paper evaluates the differences between two currently leading annotation schemes for dependency treebanks. By relying on four treebanks, we demonstrate that the treatment of conjunctions and adpositions represents the core difference between the two schemes and that this impacts the topological properties of the linguistic networks induced from the treebanks. We also show that such properties are reflected in the performances of four probabilistic dependency parsers trained on the treebanks.

Keywords: treebank, syntax, network analysis, natural language processing

1. *Introduction*

One limitation that has been affecting for years the research area that deals with developing, disseminating and exploiting syntactically annotated corpora (known as 'treebanks') is the use of different annotation schemes.

In the context of dependency treebanks, annotation schemes can differ in several aspects, ranging from the set of dependency relation labels to the treatment of specific constructions like subordinate clauses, verb groups, and coordinated and adpositional phrases[1].

These divergences represent a significant obstacle to the use of dependency treebanks in contrastive theoretical linguistics as well as in Natural Language Processing (NLP) and, particularly, in multilingual language technologies, like cross-lingual syntactic parsing[2].

An effective way to overcome such limitation is to convert the various treebanks into some common schema and to make them available in some repository. So far, two projects are attempting such task.

[1] D. Zeman – D. Mareček – M. Popel – L. Ramasamy – J. Štěpánek – Z. Žabokrtský – J. Hajič, *HamleDT: To parse or not to parse?*, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, N. Calzolari – K. Choukri – T. Declerck – M. Uğur Doğan – B. Maegaard – J. Mariani – A. Moreno – J. Odijk – S. Piperidis ed., European Language Resources Association (ELRA), Istanbul 2012, pp. 2735-2741.

[2] R.T. McDonald – S. Petrov – K. Hall, *Multi-source transfer of delexicalized dependency parsers*, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, R. Barzilay – M. Johnson ed., Association for Computational Linguistics, Stroudsburg, PA 2011, pp. 62-72.

The first is Universal Dependency Treebanks v2.0 (UDT v2; sometimes also referred to as Google Universal Treebanks)[3], which was published in Spring 2014, including eleven dependency treebanks in as many languages. The annotation scheme is based on Google universal part-of-speech tags[4], the Interset interlingua for morphosyntactic tagsets[5] and Universal Stanford Dependencies (USD), which adapt the previous version of the Stanford Dependencies representation to capture grammatical relations across languages[6].

The second project is HamleDT 2.0 (issued in May 2014)[7], a compilation of thirty existing dependency treebanks or dependency conversions of other treebanks. The treebanks are harmonized both into basic USD and into Prague Dependencies (PRG)[8], an annotation scheme which slightly adapts the one used in the so-called 'analytical' layer of the Prague Dependency Treebank for Czech[9].

The availability of several treebanks annotated according to USD and/or PRG makes these the currently leading and most widespread annotation schemes for dependency treebanks.

Since the empirical evidence provided by treebanks is largely used both for NLP purposes and for studies in theoretical linguistics, this paper wants to investigate (through network analysis) the differences between the two schemes and to evaluate to what extent

[3] R.T. McDonald – J. Nivre – Y. Quirmbach-Brundage – Y. Goldberg – D. Das – K. Ganchev – K. Hall – S. Petrov – H. Zang – O. Täckström – C. Bedini – N.B. Castelló – J. Lee, *Universal dependency annotation for multilingual parsing*, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, H. Schütze – P. Fung – M. Poesio ed., Association for Computational Linguistics, Stroudsburg, PA 2013, pp. 92-97. UDT v2 must not be confused with Universal Dependencies (UD), a newer project with several data releases since January 2015, http://universaldependencies.org/ (last accessed February 29, 2016).

[4] S. Petrov – D. Das – R. McDonald, *A universal part-of-speech tagset*, in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, N. Calzolari – K. Choukri – T. Declerck – M. Uğur Doğan – B. Maegaard – J. Mariani – A. Moreno – J. Odijk – S. Piperidis ed., European Language Resources Association (ELRA), Istanbul 2012, pp. 2089-2096.

[5] D. Zeman, *Reusable Tagset Conversion Using Tagset Drivers*, in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, N. Calzolari – K. Choukri – B. Maegaard – J. Mariani – J. Odijk – S. Piperidis – D. Tapias ed., European Language Resources Association (ELRA), Marrakech 2008, pp. 213-218.

[6] M.C. de Marneffe – M. Connor – N. Silveira – S.R. Bowman – T. Dozat – C.D. Manning, *More constructions, more genres: Extending Stanford dependencies*, in *DepLing 2013. Proceedings of the Second International Conference on Dependency Linguistics*, E. Hajičova – K. Gerdes – L. Wanner ed., Matfyzpress, Prague 2013, pp. 187-196. M.C. de Marneffe – N. Silveira – T. Dozat – K. Haverinen – F. Ginter – J. Nivre – C.D. Manning, *Universal Stanford dependencies: A cross-linguistic typology*, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, N. Calzolari – K. Choukri – T. Declerck – H. Loftsson – B. Maegaard – J. Mariani – A. Moreno – J. Odijk – S. Piperidis ed., European Language Resources Association (ELRA), Reykjavik 2014, pp. 4585-4592.

[7] R. Rosa – J. Mašek – D. Mareček – M. Popel – D. Zeman – Z. Žabokrtský, *HamleDT 2.0: Thirty Dependency Treebanks Stanfordized*, in *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, pp. 2334-2341.

[8] D. Zeman et alii, *HamleDT: To parse or not to parse?*.

[9] J. Hajič – J. Panevová – E. Hajičová – P. Sgall – P. Pajas – J. Štěpánek – J. Havelka – M. Mikulová – Z. Žabokrtský – M. Ševčíková-Razímová ed., *Prague Dependency Treebank 2.0,* LDC Catalog No. LDC2006T01, Philadelphia 2006.

these differences impact the overall properties of dependency treebanks and their use in research. In particular, we wonder how (and if) such properties are reflected in the performances of a number of probabilistic syntactic parsers trained and tested on treebanks available in the two schemes.

The paper is organized as follows: Section 2 describes the main differences between USD and PRG; Section 3 presents the data and evaluates the rate of similarity and difference among the treebanks for four languages annotated according to both USD and PRG; Section 4 describes and motivates the topological properties of the linguistic networks induced from the treebanks; Section 5 presents the training and testing of four probabilistic dependency parsers and discusses the results in the light of the topological properties of the networks; Section 6 presents conclusions and sketches the future work.

2. *Stanford and Prague Dependencies*

According to Rosa et alii[10], the main differences between USD and PRG are (a) the underlying theory that motivates the annotation scheme and (b) the goal itself for which the scheme has been designed.

USD build upon Lexical-Functional Grammar[11], representing a kind of dependency-based counterpart of it. Instead, the theoretical framework that motivates PRG is Functional Generative Description[12], which understands the dependency tree representing surface syntax as an intermediate (nearly technical) layer of annotation, built upon the morphological layer (which includes lemmatization and morphological tagging) and leading to the underlying syntax layer (featuring semantic role labeling, anaphora and ellipsis resolution, and annotation of information structure).

USD keep the representation of syntax as easy as possible, because the data are meant to be used in NLP applications, like stochastic parsing and information retrieval. PRG are linguistically very accurate, because they were designed to evaluate (and possibly refine) the background theory on the basis of the empirical evidence obtained while building the Prague Dependency Treebank; this may lead to quite complex representations of certain syntactic structures.

Entering USD and PRG in more detail, there are two main aspects that characterize a dependency-based annotation scheme: (a) the inventory of dependency relations used and (b) the criteria selected to design the parent-child relations between nodes in the trees (the so-called 'dependencies').

Although USD and PRG use different sets of dependency relations, some of them can be converted from one scheme into the other quite regularly. For instance, the 'mark' rela-

[10] R. Rosa et alii, *HamleDT 2.0: Thirty Dependency Treebanks Stanfordized.*

[11] J. Bresnan, *Lexical-functional syntax*, Wiley-Blackwell, Oxford 2001.

[12] P. Sgall – E. Hajicová – J. Panevová, *The meaning of the sentence in its semantic and pragmatic aspects*, Reidel, Dordrecht 1986.

tion in USD (assigned to subordinating conjunctions) corresponds to the 'AuxC' relation in PRG without exceptions.

In both USD and PRG, the criteria for assigning parent-child relations observe the basic principles of dependency grammar, like for instance the head role assigned to predicates and the dependence of attributes on nouns. However, a number of differences between the two schemes do hold when specific constructions are concerned. In particular, it is well-known that USD and PRG differ in the way they treat copular constructions, (subordinating and coordinating) conjunctions and adpositions (i.e. prepositions and postpositions):

- in USD, the nominal predicate in copular constructions governs the copula, while the opposite holds in PRG;
- in USD, adpositions and subordinating conjunctions are governed by the word they introduce; in PRG, both adpositions and subordinating conjunctions govern the head of their respective phrases, acting as auxiliary elements that bridge the heads of two phrases standing in parent-child relation;
- in USD, the conjuncts in coordination constructions are siblings except for the first one, which heads the other conjunct(s) and the coordinating conjunction(s). Instead, in PRG, the coordinating conjunction (or a punctuation fulfilling its role) governs the conjuncts, which are all siblings and assigned a specific extension (_M)[13]. This difference in treating coordination implies that in USD the first conjunct is not labeled as a conjunct explicitly, but this can be deduced only from the presence of conjuncts among its children. Furthermore, while USD does not distinguish between private and shared modifiers (because this cannot be done topologically), in PRG this is marked by the absence of the extension _M in the label assigned to one or more children of the coordinating conjunction.

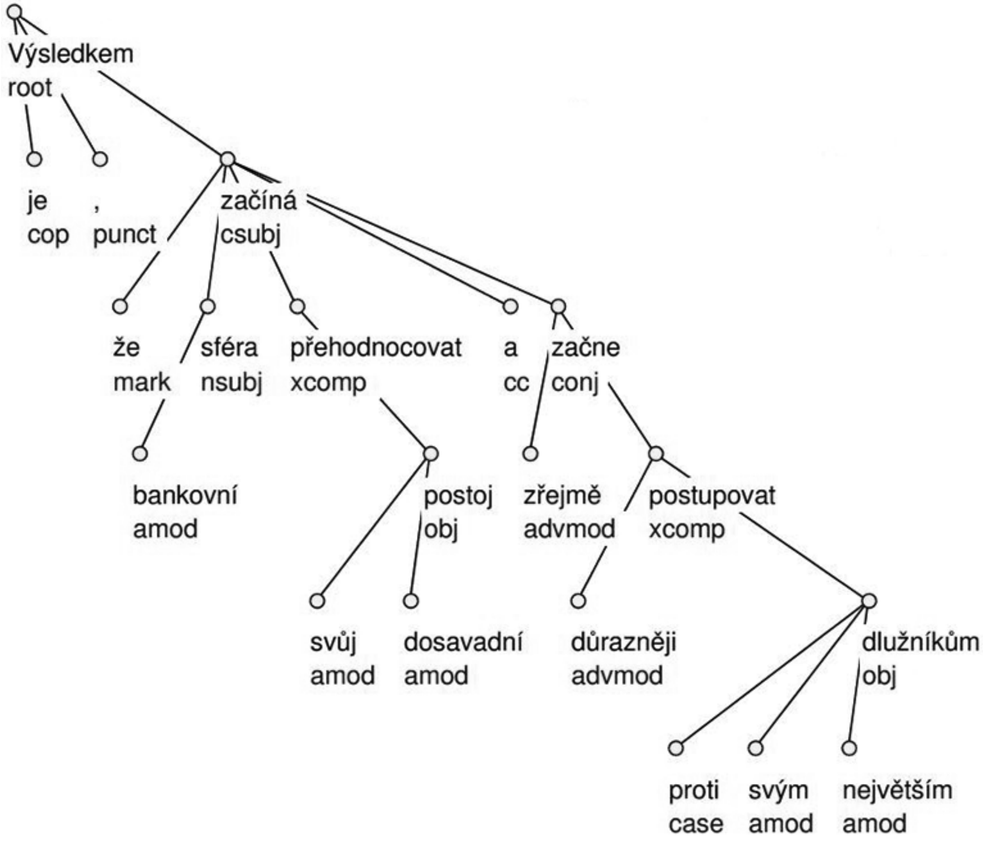
Figures 1 and 2 present respectively the USD tree and the PRG tree for sentence (1) taken from the treebanks for Czech provided by HamleDT 2.0.

(1)
Výsledkem [Result] je [is], že [that] bankovní [banking] sféra [sector] začíná [begins] přehodnocovat [to re-evaluate] svůj [its] dosavadní [current] postoj [position] a [and] zřejmě [likely] začne [begins] důrazněji [more forcefully] postupovat [to act] proti [against] svým [their] největším [biggest] dlužníkům [debtors]

The result is that the banking sector begins to re-evaluate its current position and will likely act more forcefully against their biggest debtors

[13] In PRG, if more than one coordinating conjunction is present (multiple coordination), the rightmost conjunction in the text governs the other(s) (the leftmost in right-to-left languages).

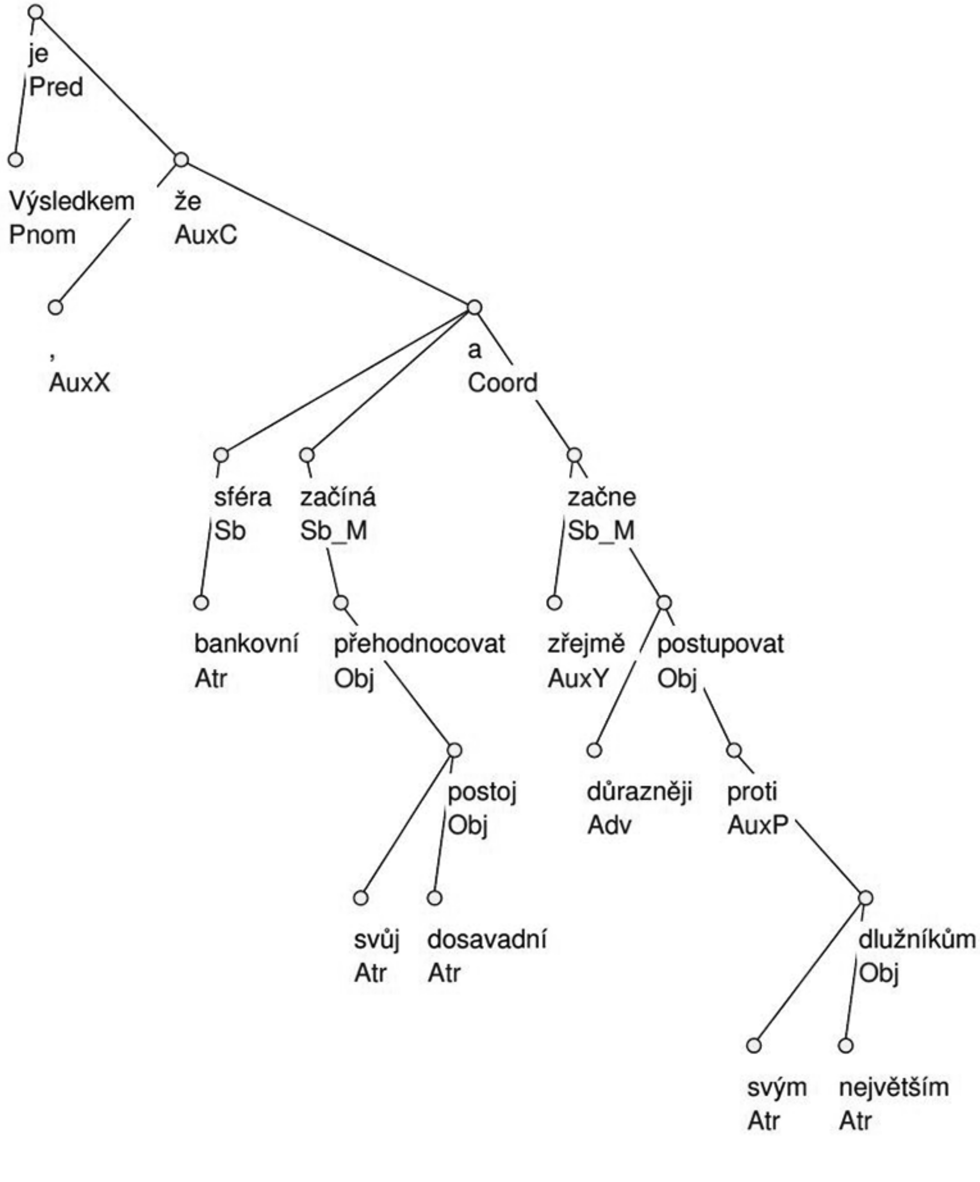
Figure 1 - *A USD tree from HamleDT 2.0*



The dependency trees in figures 1 and 2 differ in the treatment of the following nodes/constructions:

- copular constructions. USD: the copula *je* depends on the nominal predicate (*Výsledkem*). PRG: the opposite;
- adpositions. USD: the preposition *proti* depends on the noun that it modifies (*dlužníkům*). PRG: the opposite;
- subordinating conjunctions. USD: the subordinating conjunction *že* depends on the first predicate of the coordinated subordinate clause it introduces (*začíná*). PRG: *že* depends on the head node of the governing clause (*je*);
- coordinating conjunctions. USD: the coordinating conjunction *a* depends on the first of the two conjuncts (*začíná*). PRG: *a* governs both the conjuncts (*začíná* and *začne*);
- shared modifiers. USD: the shared modifier *sféra* depends on the first of the two conjuncts (*začíná*). PRG: *sféra* depends on the coordinating conjunction (*a*) and the absence of the extension _M in its label informs that it is a modifier shared by all the conjuncts (*začíná* and *začne*).

Figure 2 - *A PRG tree from HamleDT 2.0*



In HamleDT 2.0, the treebanks were first harmonized into PRG and then 'stanfordized' by rehanging some of the nodes in the trees and mapping the PRG labels to USD ones. In particular, the PRG representation of coordinating structures has been demonstrated to have more expressive power than USD[14]; thus, converting these structures from PRG to

[14] M. Popel – D. Marecek – J. Stepánek – D. Zeman – Z. Zabokrtský, *Coordination Structures in Dependency Treebanks*. in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 517-527.

USD did not raise particular problems. Since USD distinguish direct and indirect objects (by using the 'dobj' and 'iobj' labels, respectively), which PRG do not, an 'obj' label was added in the tagset in place of 'dobj' and 'iobj'.

3. *Comparing Treebanks and Schemes*

In order to compare the two annotation schemes, we first evaluated the degree of similarity of the same data annotated both in USD and in PRG.

We selected four out of the thirteen treebanks made available by HamleDT 2.0. These are the Prague Dependency Treebank for Czech[15], the Alpino Dependency Treebank for Dutch[16], the Persian Dependency Treebank for Persian[17] and the Floresta Sintá(c)tica treebank for Portuguese[18].

We chose these treebanks because (a) they provide evidence about languages belonging to different linguistic groups (Slavic, Germanic, Indo-Iranian and Romance, respectively) and (b) they are the largest ones among those provided with the most free license in HamleDT 2.0, which is needed for training and testing probabilistic NLP tools (see section 5)[19].

To overcome the differences in size among the treebanks, we selected the first 150,000 nodes from each treebank[20]. Then, we compared the treebanks for each language in the two annotation schemes, by calculating the percentage of nodes that share the same parent node in the two treebanks, regardless of the dependency relation. We used this metric because the sets of dependency relations of USD and PRG are completely different. Thus, dependency relations do not provide any efficient hint to compare the treebanks, resulting

[15] E. Bejček – E. Hajičová – J. Hajič – P. Jínová – V. Kettnerová – V. Kolářová – M. Mikulová – J. Mírovský – A. Nedoluzhko – J. Panevová – L. Poláková – M. Ševčíková – J. Štěpánek – Š. Zikánová ed., *Prague Dependency Treebank 3.0*, Charles University in Prague, ÚFAL, Prague 2013. http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3 (last accessed February 29, 2016).

[16] L. van der Beek – G. Bouma – R. Malouf – G. van Noord, *The Alpino Dependency Treebank*, "Language and Computers", 45, 2002, 1, pp. 8-22.

[17] M. Sadegh Rasooli – A. Moloodi – M. Kouhestani – B. Minaei-Bidgoli, *A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank*, in *5th Language and Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, Z. Vetulani ed., Poznań 2011, pp. 227-231.

[18] S. Afonso – E. Bick – R. Haber – D. Santos, *"Floresta sintá(c)tica": A Treebank for Portuguese*, in *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, N. Calzolari – K. Choukri – B. Maegaard – J. Mariani – J. Odijk – S. Piperidis – D. Tapias ed., European Language Resources Association (ELRA), Las Palmas 2002, pp. 1698-1703.

[19] Number of nodes in the 'train' set of the treebanks: Czech: 331,242; Dutch: 195,069; Persian: 182,878; Portuguese: 331,242. Although also the treebank for Arabic provided by HamleDT 2.0 is large enough (249,600 nodes), we could not use it because it includes hundreds of sentences with more than 100 nodes, which are too long for running graph-based parsers on them at both training and testing level with the machines at our disposal (see section 5).

[20] We selected the data by sentence boundary, cutting the treebanks at the first end of the sentence after node n. 150,000. This resulted in the following number of nodes in the single treebanks: Czech: 150,012; Dutch: 150,013; Persian: 150,009; Portuguese: 150,008.

in zero similarity. Instead, the parent-child relations are partly shared by the two annotation schemes: calculating how many of them are the same in the two treebanks for each language is an efficient way to evaluate the degree of similarity of the treebanks and, more generally, of the two annotation schemes.

Table 1 presents the percentage of nodes with the same parent node in both USD and PRG treebanks for each language, regardless of the dependency relation.

Table 1 - *Similarity of USD and PRG treebanks by parent-child relations*

Czech	*Dutch*	*Persian*	*Portuguese*
0.525	0.524	0.307	0.525

The results reported in table 1 show that USD and PRG treebanks share slightly more than half of the dependencies. The results are very similar for all languages, Persian representing the only exception, with a lower percentage (0.307).

Although table 1 informs about the general rate of similarity of the treebanks in the two annotation schemes, it fails to provide any specific insight about which particular dependencies are the same, and which are not, in the two schemes. In order to detail this, we calculated the percentage of nodes with the same parent node in the treebanks by part-of-speech (PoS). Table 2 shows the results[21].

Table 2 - *Similarity of USD and PRG treebanks by PoS-based parent-child relations*

	Czech	*Dutch*	*Persian*	*Portuguese*
Adp	0.032	0.014	0.021	0.033
Conj	0.319	0.096	0.02	0.08
Noun	0.535	0.422	0.344	0.342
Adj	0.842	0.907	0.679	0.947
Numeral	0.709	0.726	NA	0.724
Verb	0.549	0.611	0.247	0.593
Pronoun	0.81	0.787	0.604	0.848
Adverb	0.839	0.827	0.575	0.8
Punct	0.353	0.27	0.234	0.608
Particles	0.792	NA	0.586	1
Sub. Conj	0.681	0.106	0.021	0.09
Co. Conj	0.162	0.087	0.02	0.075

[21] The PoS for particles is not available in the Dutch treebanks, as well as that for numerals in the Persian treebanks.

Table 2 clearly shows that adpositions and conjunctions are the PoS with the lowest percentage of dependencies shared by the treebanks in the two schemes for all languages. Coordinating conjunctions feature a lower rate than subordinating ones.

4. *Network Analysis*

In order to better understand the different role played by adpositions and conjunctions in the two annotation schemes, and to evaluate how this impacts the overall features of the treebanks, we need some method able to inform about the general properties of USD and PRG treebanks, by providing a synoptic view and grasp of data.

Given (a) that the main formative elements of a dependency treebank are nodes and relations between them, and (b) that a network is a (un)directed graph $G(V, E)$ which is given by a set of vertices V and a set of edges E[22], representing a dependency treebank as a network whose vertices are lemmas and edges are dependencies looks like an efficient method to detect and manage the general properties of a treebank.

4.1 Building the Networks

In order to build networks from dependency treebanks, we applied the method developed by Ferrer i Cancho et alii[23]. 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 in the network if they appear at least once in a dependency relation in the treebank. In an oriented network, the edges are directed according to the direction of the dependency relation in the treebank (i.e. edges go from the parent to the child node).

Then, a syntactic dependency network is built 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 the connections. The network is an emergent property of sentence structures[24], while the structure of a single sentence is a subgraph of the global network[25].

We applied this method to build the corresponding oriented syntactic dependency network from each treebank used in this work. In total, we built eight networks from eight

[22] R. Ferrer i Cancho, *Network theory*, in *The Cambridge Encyclopedia of the Language Sciences*, P. Colm Hogan ed., Cambridge University Press, Cambridge, UK 2010, pp. 555-557.

[23] R. Ferrer i Cancho – R.V. Solé – R. Köhler, *Patterns in syntactic dependency networks*, "Physical Review", E69, 2004, 051915(8).

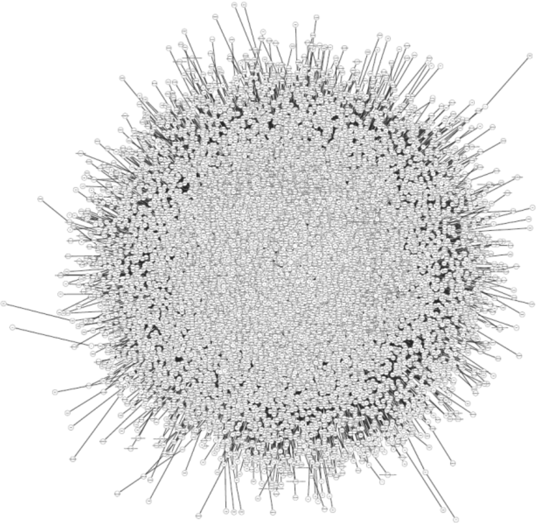
[24] R. Ferrer i Cancho, *The structure of syntactic dependency networks: insights from recent advances in network theory*, in *Problems of quantitative linguistics*, G. Altmann – V. Levickij – V. Perebyinis ed., RAM-Verlag, Lüdenscheid 2005, pp. 60-75.

[25] B. Bollobás, *Modern Graph Theory*, Springer, New York 1998 (Graduate Texts in Mathematics, 184).

treebanks (four languages, each one in two annotation schemes). We used the free software *Cytoscape*[26] and the R package *igraph*[27] for network creation and computing.

Figure 3 shows the syntactic dependency network built from the PRG treebank for Portuguese. Vertices and edges are arranged according to the 'Prefuse Force Directed Layout' setting provided by Cytoscape[28]. Edges are weighted by frequency, the most central relations in the network being those most frequent in the treebank.

Figure 3 - *The network of the PRG treebank for Portuguese*



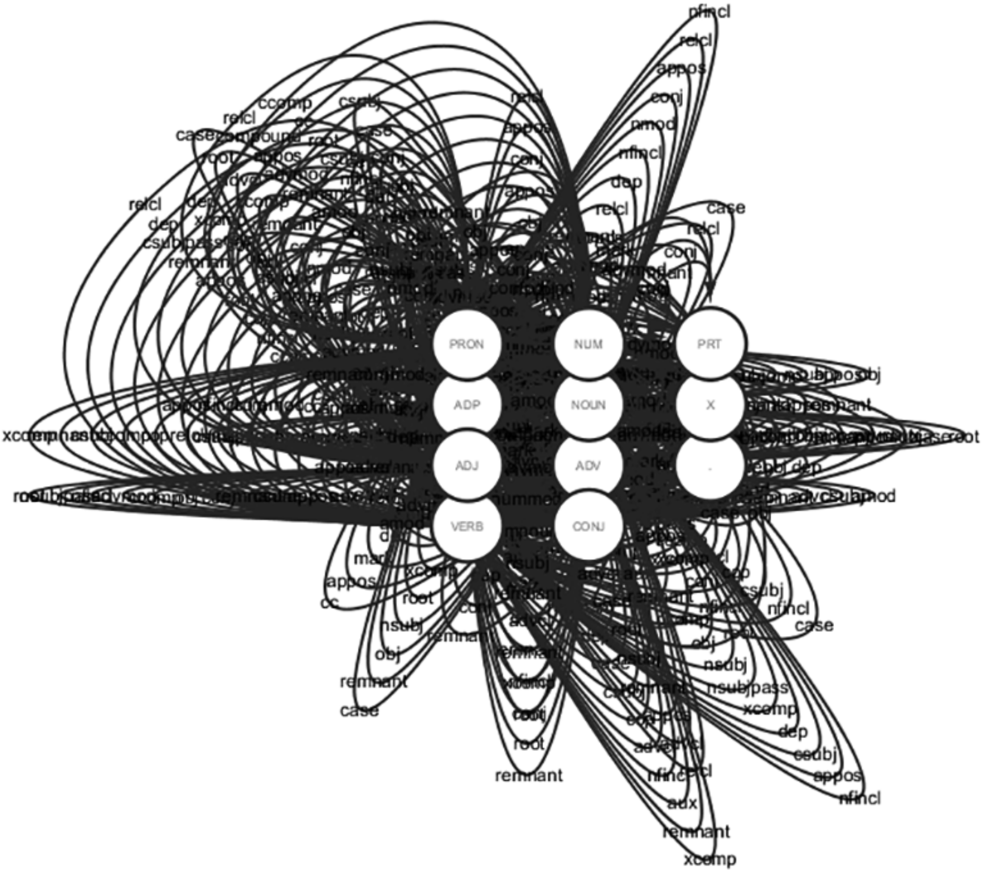
As reported above (see table 2), PoS-based parent-child relations show different degrees of similarity between USD and PRG treebanks. In order to perform network analysis of the behavior of some specific PoS in the source data, we induced the PoS-based syntactic networks from all the treebanks, thus resulting in eight PoS-based networks (like for the general networks described above). In such networks, the vertices represent single PoS (instead of lemmas) and the edges are the dependency relations holding between two PoS in the source treebank. The edges are oriented from parent to child, counted (reflecting the frequency of the connections in the source treebank) and labelled with syntactic relations. For instance, figure 4 shows the PoS-based network built from the USD treebank for Czech.

[26] R. Saito – M.E. Smoot – K. Ono – J. Ruscheinski – P.L. Wang – S. Lotia – A.R. Pico – G.D. Bader – T. Ideker, *A travel guide to Cytoscape plugins*, "Nature Methods", 9, 2012, 11, pp. 1069-1076.

[27] G. Csardi – T. Nepusz, *The igraph software package for complex network research*, "InterJournal, Complex Systems", 1695, 2006, 5, pp. 1-9.

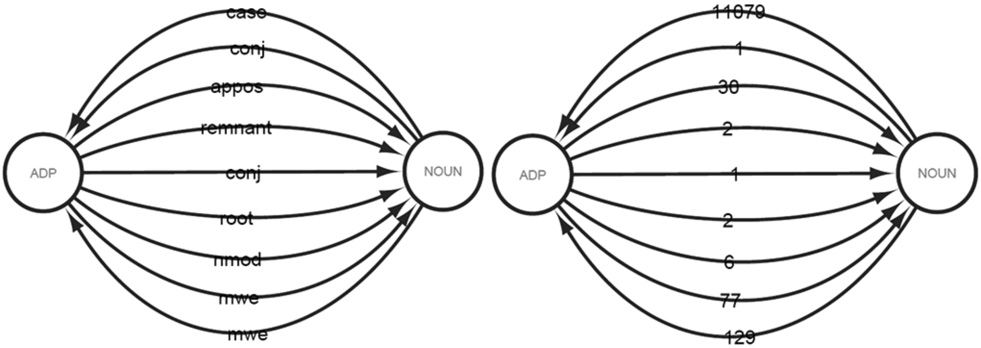
[28] M. Kohl – S. Wiese – B. Warscheid, *Cytoscape: software for visualization and analysis of biological networks*, in *Data Mining in Proteomics*, M .Hamacher – M. Eisenacher – C. Stephan ed., Humana Press, New York 2011, pp. 291-303.

Figure 4 - *The PoS-based network of the USD treebank for Czech*



In order to further clarify the structure of a PoS-based network, figure 5 represents a sub-network of figure 4. In particular, figure 5 shows the vertices for two PoS, namely adpositions (ADP) and nouns (NOUN), and the edges holding between them. The latter are directed and labelled respectively with syntactic relations in left figure 5 and with frequencies in right figure 5. For instance, the top edge appearing in left figure 5 goes from the NOUN vertex to the ADP one and it is labelled with the syntactic relation ‘case’, which in USD labels the case-marking elements treated as a separate syntactic word (like adpositions and clitic case markers). This means that this edge represents all the dependencies in the source treebank where a noun governs an adposition via the ‘case’ relation. Given that every edge in a network built from a dependency treebank is assigned the frequency of the connection that it represents, right figure 5 informs that ‘case’ is the most frequent relation holding between nouns and adpositions in the USD treebank for Czech (11,079 occurrences).

Figure 5 - *A subnetwork of the PoS-based network of the USD treebank for Czech*



The drawings in figures 3 and 4 are messy and not very informative. In order to both analyze and categorize the networks, we used a number of topological indices that are able to unravel fundamental structural properties of the networks that are hidden to the eye.

4.2 Analyzing the Networks: Assortativity

(Linguistic) networks can be analyzed through several topological indices[29], which inform about various structural properties of the networks.

In order to analyze the linguistic networks that we built from our set of treebanks, we first used a topological index called 'assortativity'. Assortativity is a property of networks that describes connectivity preferences among vertices. Roughly speaking, assortativity informs whether in a network vertices of degree *k* connect to vertices of degrees similar to *k* ('assortative mixing') or not ('disassortative mixing')[30].

Assortative mixing was observed for several kinds of networks, like for instance social networks[31]. Disassortative mixing was shown for Wiki and document networks[32] as well as for syntactic dependency networks[33].

Disassortative mixing is typical of linguistic networks, because they feature many vertices with a few connections and a few vertices with a disproportionately large number of connections. Among the connections of the vertices of the latter type are both vertices

[29] O. Abramov – A. Mehler, *Automatic Language Classification by means of Syntactic Dependency Networks*, "Journal of Quantitative Linguistics", 18, 2011, 4, pp. 291-336.

[30] The degree of a vertex *s* is the number of its edges, i.e. different relations holding between *s* and other vertices in the network. In a linguistic network, the degree of a vertex (i.e. a lemma) is strictly, although not directly, related to the frequency of that lemma in the input data. In an oriented network, the degree results from the sum of the out-degree, which labels the number of edges that are directed from the vertex, and of the in-degree, which labels the number of edges that are directed to the vertex.

[31] M.E.J. Newman – J. Park, *Why social networks are different from other types of networks*, "Physical Review", E68, 2003, 036122(3).

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

[33] O. Abramov – A. Mehler, *Automatic Language Classification*.

with high degree and vertices with low degree, which is a sign of the disassortativity of a network.

Table 3 shows the values of assortativity for the eight networks that we built.

Table 3 - *(Dis)assortativity*

	Czech	*Dutch*	*Persian*	*Portuguese*
USD	–0.126	–0.201	–0.176	–0.183
PRG	–0.175	–0.238	–0.235	–0.242

Not surprisingly, all the networks show negative values of assortativity, which means that they all present disassortative mixing. More in detail, the networks built from PRG treebanks are always more disassortative than the corresponding USD ones. This result can be explained by the differences in the treatment of some dependencies in the two annotation schemes and, in particular, by that of adpositions and conjunctions, which are the PoS with the lowest percentage of dependencies common to USD and PRG treebanks (see section 3).

As said, adpositions and conjunctions act like bridge-nodes in PRG, connecting the heads of two phrases. Instead, in USD they depend on the head of their phrase (or on the first conjunct, in the case of coordinating conjunctions). This results in a generally lower degree of the vertices for adposition and conjunctions in the USD networks than in the corresponding PRG ones. Indeed, while in a USD treebank adpositions and conjunctions are (usually) connected to one node only (which they depend on), in a PRG treebank they are (usually) connected to two nodes, i.e. one parent and one – and possibly more than one – child.

Since both adpositions and conjunctions are highly frequent PoS in the treebanks, they have high degree in the networks. Assortativity is a topological index that evaluates if the vertices of a network are connected to vertices of similar degree or not. Both adpositions and conjunctions show a heterogeneous distribution of connections: this means that they are connected to vertices with a wide range of degrees, but mostly to vertices of low degree[34]. Thus, if conjunctions and adpositions have higher degree in a network, this results in higher disassortative mixing for that network, just because there is a higher number of vertices of low degree that are connected to vertices of high degree (i.e. those for adpositions and conjunctions).

For example, let's consider the lemma *de* [of], which is the most frequent preposition in the treebanks for Portuguese. Table 4 shows the number of connections of this vertex in the USD and in the PRG networks built from the Portuguese treebanks.

Consistently with the bridging role played by adpositions in the PRG scheme, the vertex for *de* is much more connected in the PRG network than in the USD one: its degree is higher in PRG (7,672 vs. 4,603), as also the number of different vertices which it is connected to (6,068 vs. 4,536).

[34] For Zipf's law, a text features a few words with very high frequency and a large number of words with low frequency. See G.K. Zipf, *Human behavior and the principle of least effort*, Addison-Wesley, Reading 1949.

Furthermore, the out-degree of *de* in the USD network is dramatically lower than its in-degree (103 vs. 4,500), which results from the fact that in the USD scheme adpositions mostly act as child nodes and almost never as parent nodes. Things are clearly different in the PRG network, where the values for the out-degree and the in-degree of *de* are much closer than in the USD one (4,528 vs. 3,144).

Table 4 - *Connectivity of* de[35]

	Degree	*In-degree*	*Out-degree*	*Vertices*
USD	4,603	4,500	103	4,536
PRG	7,672	3,144	4,528	6,068

As mentioned, higher disassortative mixing results from higher number of connections of a vertex of degree *k* with vertices of degree (very) different from *k*. In this respect, adpositions contribute heavily to make a network disassortative, because most of the vertices which they are connected to show a degree much lower than them. In the two networks for Portuguese, we calculated the number of vertices that are directly connected to that of *de* and have a degree higher than the 10% of the degree of *de* (i.e. higher than 460 in USD and higher than 767 in PRG). Among the direct connections of *de*, such vertices are the ones that least contribute to improve the disassortativity of the network, because they are those with the 'less different' degree from *de*. Thus, the lower is the number of such vertices, the higher is the disassortative mixing of the network, and vice versa.

Such vertices are 21 in the USD network and 16 in the PRG network, corresponding respectively to 0.26% of the vertices directly connected to *de* in the PRG network (16/6,068) and to 0.46% in the USD network (21/4,536). Although both percentages are very low (thus, confirming the highly heterogeneous and disassortative connectivity of *de*), the USD value is almost double than the PRG one, which explains the higher disassortative mixing of the PRG network in comparison to the USD one.

4.3 Analyzing the Networks: Small-worldness

The second criterion we used to analyze the linguistic networks that we built from the treebanks is their degree of 'small-worldness'.

The degree of small-worldness of a network is related to its connectedness, or compactness. The term 'small-world' comes from the observation in the social sciences that everyone in the world can be reached through a short chain of social acquaintances although the number of people of the whole social network is huge. Networks of different kind tend to be small worlds. Despite the large amount of vertices in networks, the distance between them is surprisingly small. This means that, regardless of

[35] In table 4, the column 'Vertices' reports the number of vertices to which the vertex for *de* is directly connected in the network (i.e. its 'connections'). The total of connections of *de* is lower than its degree, because one vertex can be connected to that for *de* by more than one edge (maximum two edges: one entering and one exiting *de*).

their dimension, networks tend to be highly compact and very well connected: it is very easy to reach a given element from another one through a small number of jumps.

According to the so-called 'Small-World Model' by Watts and Strogatz[36], a network is said to be a small world if it shows low average shortest path length and high clustering coefficient.

Path length is defined as the average minimal distance between any pair of vertices[37]. The 'average shortest path length' is defined as the average shortest distance between any pair of vertices in a network.

'Clustering coefficient' is the probability that two vertices that are neighbors of a given vertex are neighbors of each other[38]. In other words, it is a measure of the relative frequency of triangles in a network.

Tables 5 and 6 show respectively the values for average shortest path length and clustering coefficient resulting from the USD and PRG networks of the four languages here concerned.

Table 5 - *Average shortest path lengths*

	Czech	*Dutch*	*Persian*	*Portuguese*
USD	4.125	4.018	3.726	3.848
PRG	3.487	3.738	3.46	3.124

Table 6 - *Clustering coefficients*

	Czech	*Dutch*	*Persian*	*Portuguese*
USD	0.079	0.106	0.097	0.176
PRG	0.146	0.132	0.191	0.312

All the PRG networks have lower average shortest path length and higher clustering coefficient than the corresponding USD networks. This means that PRG networks are more small-world than USD ones. Again, adpositions and conjunctions help to explain this.

Following their high frequency in data, adpositions and conjunctions are among the most connected vertices in linguistic networks. Such vertices are called 'hubs'[39]. Hubs are the key components of the complexity of a network, supporting high efficiency of network traversal. Just because of such an important role in the network, their loss heavily impacts the performance of the whole system, whose properties change radically[40]. For instance, re-

[36] D.J. Watts – S.H. Strogatz, *Collective dynamics of 'small-world' networks*, "Nature", 393, 1998, pp. 440-442.

[37] R.V. Solé – B. Corominas-Murtra – S. Valverde – L. Steels, *Language networks: Their structure, function, and evolution*, "Complexity", 15, 2010, 6, pp. 20-26.

[38] *Ibidem*.

[39] M.E.J. Newman, *The Structure and Function of Complex Networks*, "SIAM Review", 45, 2003, 2, pp. 167-256.

[40] H. Jeong – S.P. Mason – A.L. Barabási – Z.N. Oltvai, *Lethality and Centrality in Protein Networks*, "Nature", 411, 2001, pp. 41-42. R. Albert – H. Jeong – A.L. Barabási, *Error and attack tolerance of complex networks*,

moving from the Czech networks the vertices for just the two most frequent prepositions (*v* [in] and *na* [on]) and conjunctions (*a* [and] and *že* [that]) in the treebank, results in a decrease of the clustering coefficient and in an increase of the average shortest path length in both networks (compare table 7 with tables 5 and 6). Furthermore, a substantial amount of edges in the network gets lost[41].

Table 7 - *CCs and ASPLs for Czech networks without the two most frequent prepositions and conjunctions*

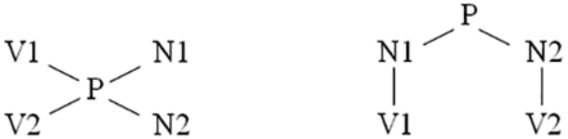
	Clustering Coefficient	*Avrg. Sh. Path Length*
USD	0.071	4.232
PRG	0.093	3.695

Due to their different treatment in the two annotation schemes, adpositions and conjunctions are 'more hubs' in the PRG networks than in the USD ones. For instance, we have seen that the preposition *de* is much more connected in the PRG network than in the corresponding USD one. But this can also be explained in more general terms.

Let's consider a preposition *P* having only two occurrences in a treebank. In both occurrences, *P* is member of a prepositional phrase (formed by the preposition itself and a noun) that modifies a verb (like, for instance, in "moving from Boston"). The nouns and the verbs (named *N1*, *N2*, *V1*, *V2*) have all different lemmas.

The dependencies among these words in the PRG scheme are the following (> means 'direct government'): V1>P>N1 and V2>P>N2. Instead, in the USD scheme, the dependencies are: V1>N1>P and V2>N2>P. Figure 6 shows the PRG network (on the left) and the USD one (on the right) corresponding to these dependencies.

Figure 6 - *PRG and USD networks for P*



In the PRG network, the vertex for *P* is equally distant (1 edge) from all the other vertices, which in turn are equally distant among themselves, by passing through *P* (2 edges): for instance, to reach *N1* from *V2* you just need to move through *P*. The degree of *P* is 4. Instead, in the USD network, the vertex for *P* is directly connected only to those for *N1* and *N2*: thus, its degree is 2.

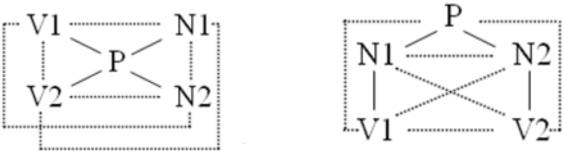
"Nature", 406, 2000, pp. 378-382.

[41] PRG: from 72,927 to 64,467. USD: from 87,801 to 83,815. The number of lost edges is higher in the PRG network just because adpositions and conjunctions are more connected in PRG networks than in USD ones.

In these networks there are ten possible paths[42]. Only two of them (*V1-N1* and *V2-N2*) are shorter in the USD network than in the PRG one (1 vs. 2); three (*N1-N2*, *P-N1* and *P-N2*) have the same length in the two networks; all the other five paths are shorter in the PRG network. For instance, the path *V1-V2* is long 2 in the PRG network and 4 in the USD one. In linguistic networks induced from real treebanks (where adpositions have very high degree), this difference in the length of the paths passing through adpositions does explode, thus explaining the higher average shortest path length for USD networks than for PRG ones.

The values for clustering coefficient can be explained as follows. If we directly connect with a dotted edge all the vertices that are not directly connected in the two networks of figure 6 (as shown in figure 7), we build the ten possible triangles that can be obtained from these networks[43].

Figure 7 - *Fully connected PRG and USD networks for P*



While drawing the edges to directly connect all the vertices in the networks to each other, we also added those that connect the vertex for *P* with those for *V1* and *V2* in the USD network (while these edges are already present in the PRG network). Actually, such connections are very rare in USD networks, because verbs and adpositions (usually) do not stand in any direct dependency relation in the USD annotation scheme. In figure 7, five triangles out of the possible ten include *P* and *V1* and/or *V2* among their vertices: these triangles are very rare in USD networks. The same holds also for subordinating conjunctions, whose direct connections with nouns are very rare in USD networks, just because they are not supposed to stand in any direct dependency relation in USD treebanks.

This limits considerably the number of triangles that actually occur in a USD network in comparison to those present in a corresponding PRG one. Since adpositions and conjunctions are much frequent in treebanks and have high degree in linguistic networks, the probability that a triangle in a linguistic network includes a vertex for an adposition is quite high. Since the triangles featuring an adposition or a conjunction among their vertices are potentially more available in PRG networks than in USD ones, when real linguistic networks induced from large treebanks are concerned, this affects the clustering coefficient, which results higher for PRG networks than for USD ones.

[42] Namely, the possible paths are the following: *P-V1; P-V2; P-N1; P-N2; V1-V2; V1-N1; V1-N2; V2-N1; V2-N2; N1-N2.*

[43] The triangles have the following vertices: *P-N1-N2; P-V1-V2; P-V1-N1; P-V2-N2; P-V1-N2; P-N1-V2; V1-V2-N1; V1-V2-N2; V1-N1-N2; V2-N1-N2.*

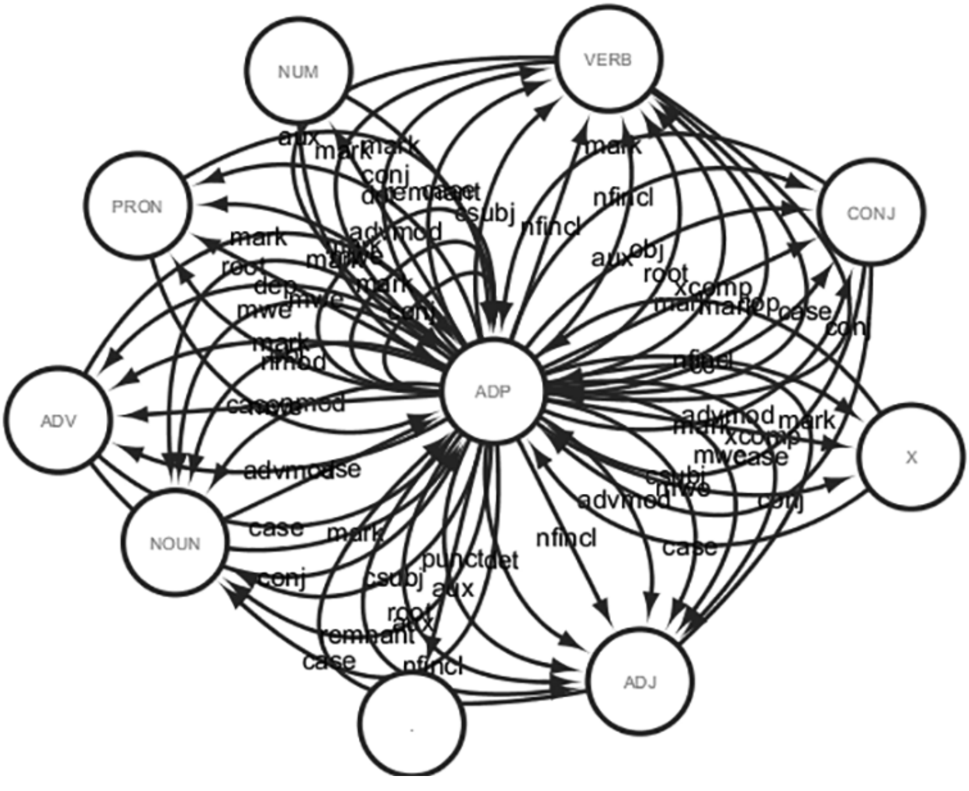
As Abramov and Mehler show[44], high disassortativity and small-world structure are topological properties typical of linguistic networks. Since PRG networks are both more disassortative and more small-world than (corresponding) USD ones, this makes PRG networks more typical linguistic networks than USD ones.

4.4 Analyzing the PoS-based Networks

While analyzing the PoS-based networks, we focused on some specific PoS. Beside adpositions and (both subordinating and coordinating) conjunctions, which are those PoS that mostly distinguish one annotation schema from the other, we also analyzed the behavior of adjectives, nouns and verbs.

For this purpose, we extracted a number of single PoS-based subnetworks from the general PoS-based networks. A single PoS-based subnetwork includes the vertex for a specific PoS, those for its direct neighbors and the edges between them (but not those holding between the neighbors themselves). For instance, figure 8 shows the adposition-based subnetwork built from the USD treebank for Dutch.

Figure 8 - *The adposition-based subnetwork of the USD treebank for Dutch*



[44] O. Abramov – A. Mehler, *Automatic Language Classification*.

We compared the single PoS-based subnetworks we built from USD and PRG treebanks by edges / vertices ratio. This ratio calculates the so-called 'average degree' of a network, i.e. the proportion of edges with respect to the number of vertices (regardless of the edge direction). Table 8 shows the average degree for the single PoS-based subnetworks we built.

Table 8 - *Edges / vertices ratio in single PoS-based subnetworks*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	PRG	USD	PRG	USD	PRG	USD	PRG	USD
Adj	11	12.1	11.2	15.2	8.07	11.9	9.22	12.44
	(110/10)	(121/10)	(112/10)	(152/10)	(113/14)	(131/11)	(83/9)	(112/9)
Adp	7.7	3.3	12.36	6.5	10.25	6	9.18	4.1
	(77/10)	(33/10)	(136/11)	(65/10)	(164/16)	(12/2)	(101/11)	(41/10)
Conj	19	8	15.27	10.6	9.87	9.64	15.7	7.6
	(190/10)	(80/10)	(168/11)	(106/10)	(158/116)	(106/11)	(157/10)	(76/10)
Noun	17	16.2	17.1	17.5	11.83	17.27	12.54	13.73
	(170/10)	(162/10)	(171/10)	(175/10)	(213/18)	(190/11)	(138/11)	(151/11)
Verb	16.1	16.09	15.54	17	10.76	12.64	16.4	15.3
	(161/10)	(177/11)	(171/11)	(170/10)	(183/17)	(139/11)	(164/10)	(153/10)

Our hypothesis is that the more similar the average degree of two single PoS-based subnetworks built from the same treebank in USD and PRG style, the more topologically similar the two subnetworks. By looking at the results, it turns out that the USD and PRG subnetworks based on conjunctions and those based on adpositions show very different average degree, while the subnetworks based on nouns, verbs and adjectives tend to present more similar average degree.

In particular, the subnetworks based on conjunctions and those on adpositions built from USD treebanks present an average degree always lower than the corresponding subnetworks built from PRG treebanks. For instance, the average degree of the conjunction-based subnetwork for Czech built from PRG treebank is 19, while that for the corresponding subnetwork built from USD treebank is 8. More in detail, the subnetworks based on conjunctions and those on adpositions built from PRG treebanks show average degree about double than those built from USD treebanks, conjunction-based subnetworks for Persian representing the only meaningful exception (PRG: 9.87; USD: 9.64).

The opposite holds for the other PoS. The average degree for the subnetworks based on adjectives, nouns and verbs tends to be lower for the subnetworks built from PRG treebanks than for those built from USD ones. Just a few exceptions to this general picture do hold. For instance, the average degree of the noun-based subnetworks for Persian is very different (PRG: 11.83; USD: 17.27). Also, the average degree of the noun-based subnetwork for Czech from the PRG treebank is slightly higher than that for the corresponding network from the USD treebank (PRG: 17; USD: 16.09).

5. *Parsing with Different Schemes*

Like other language resources, treebanks are widely used for training and testing probabilistic NLP tools. In this section we focus on a typical NLP task like dependency parsing, by wondering (a) which of the two dependency-based annotation schemes provides better parsing performances and (b) if the results are in some way related to the topological properties of the networks induced from the treebanks.

Evaluating the impact of a treebank annotation scheme on parsing results is a task that can be approached from many different perspectives, because it is an issue related to various aspects, ranging from the parsing algorithm used to the degree of granularity of the tagset and the depth of the dependency trees implied by the annotation scheme. For instance, one aspect that has attracted particular attention in this area is the different treatment of coordination structures, which has been reported to be one of the most frequent sources of parsing errors[45].

In recent years, several studies have focussed on this topic. Among them, Mille et alii[46] investigate the effect of the different degree of tagset granularity on parsing accuracy, showing that an annotation scheme provided with more fine-grained syntactic relations does not necessarily imply a significant loss in parsing accuracy. Following Kübler[47], Rehbein and van Genabith[48] evaluate a PCFG parser trained on two comparable corpora of German annotated with different schemes (TIGER and Tüba-D/Z), concluding that comparing parsing results for parsers trained on treebanks with different annotation schemes does not allow to answer the question of whether a language is harder to parse than another. Using the same two treebanks for German of Rehbein and van Genabith[49], Boyd and Meur-

[45] N. Green – Z. Žabokrtský, *Hybrid combination of constituency and dependency trees into an ensemble dependency parser*, in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, N. Grabar – M. Dupuch – A. Périnet – T. Hamon ed., Association for Computational Linguistics, Stroudsburg, PA 2012, pp. 19-26. R. McDonald – J. Nivre, *Characterizing the Errors of Data-Driven Dependency Parsing Models*, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, J. Eisner ed., Association for Computational Linguistics, Stroudsburg, PA 2007, pp. 122-131. S. Kübler – W. Maier – E. Hinrichs – E. Klett, *Parsing coordinations*, in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, A. Lascarides – C. Gardent – J. Nivre ed., Association for Computational Linguistics, Stroudsburg, PA 2009, pp. 406-414. M. Popel et alii, *Coordination Structures in Dependency Treebanks*.

[46] S. Mille – A. Burga – G. Ferraro – L. Wanner, *How Does the Granularity of an Annotation Scheme Dependency Parsing Performance?* in *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Posters,* M. Kay – C. Boitet ed., The COLING 2012 Organizing Committee, Indian Institute of Technology Bombay, Powai 2012, pp. 839-852.

[47] S. Kübler, *How Do Treebank Annotation Schemes Influence Parsing Results? Or How Not to Compare Apples And Oranges*, in *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP - 2005)*, R. Mitkov ed., Bulgarian Academy of Sciences, Borovets 2005.

[48] I. Rehbein – J. van Genabith, *Treebank Annotation Schemes and Parser Evaluation for German*, in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language*, pp. 630-639.

[49] *Ibidem*.

ers[50] demonstrate that, despite the differences in the annotation schemes, the two corpora result in comparable parsing performances. By reporting on the results of training three dependency parsers on two different Italian treebanks (TUT and ISST-TANL), Bosco et alii[51] show that, together with the peculiarities of the annotation scheme, text genre plays a significant role in affecting parsing results. Schwartz et alii[52] present a learnability-based methodology that compares pairs of annotation schemes that differ in the annotation of a single structure. The method selects the most learnable scheme, namely the one that can be best learned by a statistical parser. The authors experiment with five parsers of different types and six varying syntactic structures, showing that selecting the most learnable alternative results in higher parsing performance (with an error reduction ranging between 2.4% and 19.8%).

5.1 Results and General Evaluation

In this experiment, we used both USD and PRG treebanks to train and test four state-of-the-art probabilistic dependency parsers. In order to evaluate the impact of the two annotation schemes on different kinds of parsers, we selected two shift-reduce parsers (MaltParser v. 1.7.2[53] and DeSR v. 1.4.3[54]) and two graph-based ones (MATE-tools graph-based[55] and MSTParser v. 0.2[56]). Roughly speaking, the difference between these two methods is that shift-reduce parsers analyze sentences word by word, making decisions according to a local optimisation criterion, while graph-based parsers view sentences as a whole, making decisions according to a global criterion.

[50] A. Boyd – D. Meurers, *Revisiting the impact of different annotation schemes on PCFG parsing: a grammatical dependency evaluation*, in *Proceedings of the ACL Workshop on Parsing German (PaGe-08)*, S. Kübler – G. Penn ed., Association for Computational Linguistics, Stroudsburg, PA 2008, pp. 24-32.

[51] C. Bosco – S. Montemagni – A. Mazzei – V. Lombardo – F. Dell'Orletta – A. Lenci – L. Lesmo – G. Attardi – M. Simi – A. Lavelli – J. Hall – J. Nilsson – J. Nivre, *Comparing the influence of different treebank annotations on dependency parsing*, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, N. Calzolari – K. Choukri – B. Maegaard – J. Mariani – J. Odijk – S. Piperidis – M. Rosner – D. Tapias ed., European Language Resources Association (ELRA), Valletta 2010, pp. 1794-1801.

[52] R. Schwartz – O. Abend – A. Rappoport, *Learnability-Based Syntactic Annotation Design*, in *Proceedings of the 24th International Conference on Computational Linguistics*, pp. 2405-2422.

[53] J. Nivre – J. Hall – J. Nilsson – A. Chanev – G. Eryigit – S. Kübler – S. Marinov – E. Marsi, *MaltParser: A language-independent system for data-driven dependency parsing*, "Natural Language Engineering", 13, 2007, 2, pp. 95-135.

[54] G. Attardi – F. Dell'Orletta, *Reverse Revision and Linear Tree Combination for Dependency Parsing*, in *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT 2009). Short Papers*, M. Ostendorf – M. Collins – S. Narayanan – L. Vanderwende ed., Association for Computational Linguistics, Stroudsburg, PA 2009, pp. 261-264.

[55] B. Bohnet, *Top Accuracy and Fast Dependency Parsing is not a Contradiction*, in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, A.K. Joshi – C.R. Huang – D. Jurafsky ed., Association for Computational Linguistics, Stroudsburg, PA 2010, pp. 89-97.

[56] R. McDonald – F. Pereira, *Online Learning of Approximate Dependency Parsing Algorithms*, in *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, D. McCarthy – S. Wintner ed., Association for Computational Linguistics, Stroudsburg, PA 2006, pp. 81-88.

We trained and tested the parsers with their default settings, using a ten-fold cross validation. Tables 9 and 10 show the results by LAS and UAS[57].

Table 9 - *Parsing results on PRG data*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
DeSR	0.751	0.808	0.772	0.838	0.768	0.85	0.792	0.846
Malt	0.676	0.764	0.684	0.786	0.726	0.83	0.74	0.811
MATE	0.753	0.832	0.801	0.874	0.796	0.886	0.803	0.864
MST	0.734	0.802	0.748	0.822	0.746	0.845	0.779	0.844

Table 10 - *Parsing results on USD data*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	LAS	UAS	LAS	UAS	LAS	UAS	LAS	UAS
DeSR	0.78	0.814	0.748	0.788	0.762	0.796	0.806	0.834
Malt	0.705	0.76	0.677	0.724	0.702	0.742	0.758	0.791
MATE	0.779	0.829	0.788	0.826	0.769	0.803	0.814	0.838
MST	0.727	0.783	0.721	0.769	0.709	0.751	0.77	0.809

Actually, the LAS reported in tables 9 and 10 are not meaningful when comparing the two annotation schemes, because they are biased by the set of dependency relations used (the L in LAS). In order to understand how the differences between USD and PRG in treating some specific dependencies impact the results of probabilistic parsing regardless of the set of dependency relations, the results must be evaluated by UAS (see tables 11 and 12).

The UAS achieved by training the parsers on PRG treebanks outperform those obtained on USD ones for all languages. This holds true also for those languages whose LAS is higher on USD than on PRG (like Portuguese, with all parsers but MST). The only exception is Czech parsed with DeSR, whose UAS on USD is slightly higher than on PRG (0.814 vs. 0.808).

The gap between LAS and UAS is always much higher for PRG than for USD data. For instance, the UAS for the USD treebank for Portuguese achieved with DeSR (0.834) is less than three points higher than its LAS (0.806), but it is more than five points higher when the

[57] LAS (Labeled Attachment Score): percentage of nodes with both correct governor and dependency relation. UAS (Unlabeled Attachment Score): percentage of nodes with correct governor and wrong dependency relation. LA (Labeled Accuracy): percentage of nodes with correct dependency relation and wrong governor. See S. Buchholz – E. Marsi, *CoNLL-X Shared Task on Multilingual Dependency Parsing*, in *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, L. Màrquez – D. Klein ed., Association for Computational Linguistics, Stroudsburg, PA 2006, pp. 149-164.

PRG treebank is concerned (LAS: 0.846; UAS: 0.792). Also, the UAS for Persian obtained with MaltParser on USD data (0.742) is four points higher than its LAS (0.702), but it is more than ten points higher on the PRG treebank (LAS: 0.726; UAS: 0.83).

We wonder if such results are related to the topological properties of the networks induced from the treebanks used to train the parsers.

We have shown (see sections 4.2 and 4.3) that PRG networks are more disassortative and more small-world than the corresponding USD ones. Now we see that PRG treebanks allow higher UAS than USD ones. Our hypothesis is that the two things are connected: if we take two dependency treebanks with equal data but different dependency relations and (partly) different criteria for dependencies, the one whose network is more disassortative and more small-world tends to allow higher UAS than the other.

Such a relation between parsing performances and topological properties of linguistic networks is confirmed also if we compare the parsing performances by language. We see that Portuguese and Persian tend to provide the best (or among the best) results for both USD and PRG treebanks. The networks for these two languages (a) are the most small-world, because they present the highest clustering coefficients and the lowest average shortest path lengths (see tables 5 and 6) and (b) are among the most disassortative (see table 3). Only Dutch networks are more disassortative, but Portuguese and Persian networks show lower average shortest path length and higher clustering coefficient[58]. Indeed, interpreting the general structural properties of a network is not a matter of a single topological index, but more of a synergic overview of a number of different topological indices.

5.2 In-depth Evaluation by Single Part-of-Speech

Since adpositions and (both subordinating and coordinating) conjunctions are those PoS that mostly distinguish the USD annotation schema from the PRG one, we performed an in-depth evaluation of the parsing results on these PoS. Tables 11 and 12 show the results by UAS.

Table 11 - *UAS for adpositions*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	PRG	USD	PRG	USD	PRG	USD	PRG	USD
DeSR	0.721	0.944	0.724	0.907	0.727	0.911	0.788	0.947
Malt	0.667	0.911	0.625	0.869	0.694	0.862	0.773	0.931
MATE	0.787	0.933	0.8	0.915	0.782	0.92	0.835	0.95
MST	0.714	0.908	0.686	0.862	0.733	0.879	0.81	0.937

[58] More in detail, the PRG network for Dutch is more disassortative that the one for Persian, but less disassortative than the one for Portuguese. The USD network for Dutch is the most disassortative. Furthermore, it shows slightly higher clustering coefficient (but also much higher average shortest path length) than the Persian one.

Table 12 - *UAS for (subordinating/coordinating) conjunctions*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	PRG	USD	PRG	USD	PRG	USD	PRG	USD
DeSR	0.639	0.737	0.577	0.684	0.726	0.832	0.682	0.762
Malt	0.575	0.713	0.532	0.539	0.649	0.809	0.571	0.662
MATE	0.69	0.78	0714	0.74	0.811	0.826	0.715	0.765
MST	0.609	0.745	0.608	0.674	0.708	0.796	0.659	0.727

For both adpositions and conjunctions, the accuracy rates achieved on USD treebanks (for all languages) are always higher than those on PRG ones. In particular, the gap between the scores achieved from the treebanks in the two annotation schemes tends to be larger for adpositions than for conjunctions.

Things are different if we focus on adjectives, nouns and verbs. The accuracy rates on these PoS are very similar for all the parsers used in the experiment. For instance, table 13 shows the UAS for adjectives, nouns and verbs achieved with MaltParser.

Table 13 - *UAS for adjectives, nouns and verbs (MaltParser)*

	Czech		*Dutch*		*Persian*		*Portuguese*	
	PRG	USD	PRG	USD	PRG	USD	PRG	USD
Adj	0.889	0.897	0.904	0.901	0.832	0.78	0.965	0.965
Noun	0.687	0.62	0.705	0.649	0.661	0.562	0.754	0.693
Verb	0.639	0.694	0.812	0.745	0.632	0.619	0.613	0.631

The UAS reported in table 13 are very similar for all languages in both the annotation schemes. For instance, the UAS for adjectives achieved from the PRG treebank for Dutch is 0.904, while that from the USD treebank for the same language is 0.901. Only a few exceptions do hold, like for instance the UAS for nouns in Persian treebanks (PRG: 0.661; USD: 0.562).

On these PoS, PRG treebanks tend to perform better than USD ones[59]. This is different from what happens for conjunctions and adpositions, where the opposite case holds.

To sum up, the UAS for adpositions and conjunctions are very different for PRG and USD treebanks, the latter allowing higher results than the former. Instead, the UAS for adjectives, nouns and verbs are quite similar for the treebanks in the two annotation schemes, with the tendency of PRG treebanks to allow slightly better results than USD ones.

Now, let's compare the parsing accuracy rates with the results of the analysis performed on the single PoS-based networks (see section 4.4). Here, we see two main aspects.

[59] Again, this is valid for all the parsers that we used, although here we report only the results achieved with MaltParser.

First, the average degree of the subnetworks for conjunctions and adpositions built from PRG treebanks is much different from the average degree of those built from USD treebanks. Likewise, for what concerns parsing, the UAS for conjunctions and adpositions from PRG treebanks is much different from that from USD treebanks. Conversely, the average degree of the subnetworks for adjectives, nouns and verbs built from PRG treebanks is much similar to the average degree of those built from USD treebanks. Likewise, the UAS for adjectives, nouns and verbs from PRG treebanks tends to be similar to that from USD treebanks.

Second, the average degree of the subnetworks for conjunctions and adpositions built from USD treebanks is always lower than the average degree of those built from PRG treebanks. For what concerns parsing, the UAS for conjunctions and adpositions from USD treebanks is higher than that from PRG treebanks. Conversely, the average degree of the subnetworks for adjectives, nouns and verbs built from PRG treebanks tends to be lower than the average degree of those built from USD treebanks. For what concerns parsing, the UAS for adjectives, nouns and verbs from PRG treebanks tends to be higher than that from USD treebanks.

From these observations, we conclude the following general tendencies:

a. the more similar/different is the average degree of two single PoS-based subnetworks induced from two dependency treebanks built from the same data and annotated according to USD and PRG scheme respectively, the more similar/different are the parsing results on that PoS from the two treebanks;

b. the lower/higher is the average degree of two single PoS-based subnetworks induced from two dependency treebanks built from the same data and annotated according to USD and PRG scheme respectively, the higher/lower are the parsing results on that PoS from the two treebanks.

The A statement connects the average degree of two single PoS-based syntactic subnetworks built from two source treebanks with the degree of similarity of the parsing accuracy rates achieved on that PoS from those treebanks. The subnetwork for a specific PoS reflects the behavior of that PoS in the source treebank in terms of relations (edges) with the other PoS (vertices). If two single PoS-based subnetworks induced from the same treebank annotated in two different styles have similar topological properties, this means that the PoS in question ‘behaves similarly’ in the two treebanks with regard to its relations with the other PoS (regardless of the annotation style of the treebank).

The B statement goes one step further, by connecting the degree of complexity of a single PoS-based subnetwork with parsing results. The degree of complexity of the subnetwork is given by the average degree. If the average degree is lower, this means that the PoS in question shows a lower degree of complexity in terms of relations with the other PoS in the source treebank, which results in higher parsing accuracy. Conversely, if the average degree is higher, the PoS in question shows a higher degree of complexity in the source treebank, thus resulting in lower parsing accuracy

6. *Conclusion*

The treatment of conjunctions and adpositions is one of the aspects that mostly makes dependency-based annotation schemes different from each other.

By comparing two of the most widespread annotation schemes for dependency treebanks (USD and PRG) and exploiting a collection of treebanks for four different languages annotated in the two schemes, this paper provides an in-depth understanding of their similarities and differences.

Our results demonstrate that USD and PRG treebanks tend to have in common around half of the dependencies and that it is the different treatment of adpositions and conjunctions that mostly affects the not-shared dependencies.

While looking at the treebanks in a synoptic fashion through network analysis, we highlight the consequences that the different annotation schemes have on the overall structure of the annotated data. Furthermore, we have shown that some properties of the linguistic networks induced from the treebanks are reflected in the performances of four probabilistic dependency parsers trained on the same treebanks. In this respect, our work is just a first attempt. More treebanks in more languages must be analyzed. Also, more fine-grained settings for parsers must be tested to confirm that the theoretically explicable topological properties of linguistic networks are indeed related to the performances of probabilistic syntactic parsers.

Acknowledgements

Special thanks should be given to Felice Dell'Orletta (ILC-CNR, Pisa, Italy) and Edoardo Maria Ponti (University of Pavia, Italy) for their help in training, testing and evaluating the parsers.

FACOLTÀ DI SCIENZE LINGUISTICHE E LETTERATURE STRANIERE

L'ANALISI LINGUISTICA E LETTERARIA

ANNO XXIV - 1/2016

EDUCatt - Ente per il Diritto allo Studio Universitario dell'Università Cattolica
Largo Gemelli 1, 20123 Milano - tel. 02.72342235 - fax 02.80.53.215
e-mail: editoriale.dsu@educatt.it (produzione)
librario.dsu@educatt.it (distribuzione)
redazione.all@unicatt.it (Redazione della Rivista)
web: www.educatt.it/libri/all

ISSN 1122 - 1917

9 788893 350587