Dissecting Treebanks to Uncover Typological Trends. A Multilingual Comparative Approach

Chiara Alzetta[•◇], **Felice Dell'Orletta**[◇], **Simonetta Montemagni**[◇], **Giulia Venturi**[◇]
[•]DIBRIS, Università degli Studi di Genova, Italy
chiara.alzetta@edu.unige.it
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC-CNR) - ItaliaNLP Lab
{name.surname}@ilc.cnr.it

Introduction and Motivation. Over the last years, linguistic typology started attracting the interest of the community working on cross- and multi-lingual NLP as a way to tackle the bottleneck deriving from the lack of annotated data for many languages. Typological information is mostly acquired from publicly accessible typological databases, manually constructed by linguists. As reported in Ponti et al. (2018), despite the abundant information contained in them for many languages, these resources suffer from two main shortcomings, i.e. their limited coverage and the discrete nature of features (only "the majority value rather than the full range of possible values and their corresponding frequencies" is reported). Corpus-based studies can help to automatically acquire quantitative typological evidence which might be exploited for polyglot NLP. Recently, the availability of corpora annotated following a cross-linguistically consistent annotation scheme such as the one developed in the Universal Dependencies project is prompting new comparative linguistic studies aimed to identify similarities as well as idiosyncrasies among typologically different languages (Nivre, 2015). The line of research described here is aimed at acquiring quantitative typological evidence from UD treebanks through a multilingual contrastive approach.

Method. The proposed methodology is inspired by Alzetta et al. (2018) where an algorithm originally developed for assessing the plausibility of automatically produced syntactic annotations was used to infer quantitative typological evidence from treebanks. The authors demonstrate that the linguistic properties used by this algorithm to rank dependency annotations from reliable to unreliable ones can also be effectively used against manually revised corpora, i.e. gold treebanks. In this case, the resulting ranking of gold dependencies turned out to closely reflect the degree of prototypicality of dependency relations in the target corpus. In this study, we rely on the same algorithm, LISCA (Dell'Orletta et al., 2013), which operates in two steps: 1) it collects statistics about a set of linguistically-motivated features considering the overall tree structure extracted from a large reference corpus of automatically parsed sentences (see Figure 1), and 2) it uses them to assign a score to each dependency arc contained in a target corpus. Rather than the plausibility of the annotation, the score should be seen here as reflecting the prototypicality degree of a given relation, based on wide variety of features including its context of occurrence. The higher the score of a ranked arc, the more prototypical is the arc with respect to the statistics acquired from the large reference corpus. In Alzetta et al. (2018), the algorithm was used to acquire typological evidence from treebanks relying on LISCA models (LM) of the same language. The main novelty of this study consists in the adopted multilingual comparative approach through which typological evidence is acquired. As illustrated in Figure 2, we ranked the same monolingual treebank using LMs built for different languages, thus obtaining four different dependencies rankings of the same monolingual treebank. Different positions of the same depen-

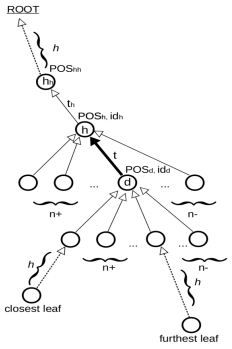

Figure 1: Features used by LISCA.

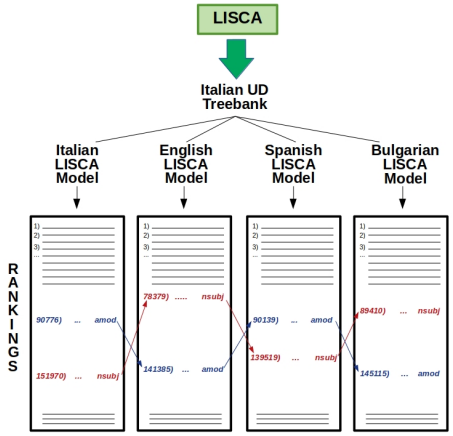

Figure 2: Method work–flow exemplified on IUDT.

LM	Target UDT models			
	IT	EN	SP	BUL
IT	**1.00**	0.94	0.97	0.79
EN	0.84	**1.00**	0.87	0.90
SP	0.98	0.95	**1.00**	0.93
BUL	0.79	0.91	0.83	**1.00**

Table 1: Spearman's correlation between pairs of ranked *amod* using different LMs on each treebank.

dency relation (DR) across the four rankings reflect different degrees of prototypicality of that DR instance: a bigger ranking difference associated with the same DR is connected with stronger typological differences of the languages represented in the selected treebanks, whereas closer rankings reflect typological closeness of languages.

Data. In this study, we considered four UD treebanks (v2.2) (Nivre et al., 2017): English (Silveira et al., 2014), Italian (Bosco et al., 2013), Spanish (McDonald et al., 2013) and Bulgarian (Simov et al., 2005). Statistics to build the LISCA models for the examined languages were extracted from four monolingual corpora of around 40 million tokens each morpho-syntactically annotated and parsed by the UDPipe pipeline (Straka et al., 2016) trained on the UD treebanks.

Results. Due to space constraints, the methodology is illustrated here wrt Italian UD Treebank (IUDT) and in particular wrt an individual DR: adjectival modifier (*amod*). Table 1 reports Spearman's rank correlation coefficients (p <0.00) obtained through pairwise comparisons of *amod* DRs across the LISCA rankings. Each pair is represented by the ranking obtained using the LM of the target UD treebank (*Target UDT models*) and the ranking obtained using one of the other LMs. Interestingly, typologically similar languages such as IT and SP show higher correlation values (0.98 and 0.97 respectively) than typologically distant ones (e.g. BUL and IT). A similar trend is observed considering the dependency direction of *amod* and its syntactic head. Table 2 and Figure 3 report *i)* the average difference of positions of pre- and post-nominal adjectival modifiers in the IUDT ranking obtained using the Italian LM and the other LMs, and *ii)* the percentage distribution of fluctuations across rankings. It results that higher the number of ranking fluctuations, the more typologically distant the languages are. Namely, in the rankings obtained using LMs of EN and BUL, a higher percentage of prenominal *amod* goes up with respect to the ranking obtained using IT LM. This reflects the linguistic properties used to build LM: right-headed adjectives are more prototypical in EN and BUL than in IT, accordingly they are highly scored by LISCA. As a consequence, a higher percentage of pre-nominal *amod* goes up in the rankings obtained using the EN (86.50) and BUL (91.35) LMs wrt the ranking obtained using IT LM. In addition, the average difference of positions of prenominal adjectives going up in the rankings obtained with EN and BUL LMs is higher (21,691.18 and 30,597.76), as well as the difference of the going-down ranking fluctuations (40,934.52 and 43,003.64). This latter result reflects the lower degree of prototipicality of left-headed adjectives in EN and BUL wrt IT.

LM	Prenominal		Postnominal	
	Up	Down	Up	Down
ENG	21,691.18	2,566.57	290.13	40,934.52
SP	1,541.31	9,163.83	5,462.83	5,322.66
BUL	30,597.76	967.66	885.15	43,003.64

Table 2: Avg difference of ranking positions of IUDT pre- and post-nominal *amod* in different rankings with respect to the ranking obtained with the Italian LM.

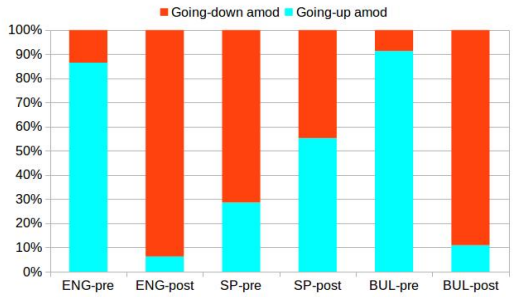

Figure 3: Distributions of the IUDT pre- and post-nominal *amod* fluctuating in different rankings with respect to the ranking obtained with the Italian LM.

References

C. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2018. Universal dependencies and quantitative typological trends. a case study on word order. In *Proceedings of the 11th Edition of International Conference on Language Resources and Evaluation (LREC 2018)*, pages 4540–4549. Association for Computational Linguistics.

C. Bosco, S. Montemagni, and M. Simi. 2013. Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria.

F. Dell'Orletta, G. Venturi, and S. Montemagni. 2013. Linguistically-driven selection of correct arcs for dependency parsing. *Computaciòn y Sistemas*, 2:125–136.

R. McDonald, J. Nivre, Y. Quirmbach-Brundage, Y. Goldberg, D. Das, K. Ganchev, K. Hall, S. Petrov, H. Zhang, O. Täckström, C. Bedini, N. Bertomeu Castelló, and J. Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

J. Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing - Proceedings of the 16th International Conference, CICLing 2015, Part I*, pages 3–16, Cairo, Egypt.

J. Nivre, A. Željko, A. Lars, and et alii. 2017. Universal dependencies 2.0. In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University*.

E.M. Ponti, H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen. 2018. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *arXiv preprint arXiv:1807.00914*.

N. Silveira, T. Dozat, M.C. de Marneffe, S. Bowman, M. Connor, J. Bauer, and C.D. Manning. 2014. A gold standard dependency corpus for english. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*.

K. Simov, P. Osenova, A. Simov, and M. Kouylekov. 2005. Design and Implementation of the Bulgarian HPSG-based Treebank. *Journal of Research on Language and Computation. Special Issue*, pages 495–522.

M. Straka, J. Hajic, and J. Strakova. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.