Quantitative Linguistic Investigations across Universal Dependencies Treebanks

Chiara Alzetta* ♠, **Felice Dell'Orletta***, **Simonetta Montemagni***,
Petya Osenova† ♣, **Kiril Simov**†, **Giulia Venturi***
* Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa
♠ DIBRIS, Università degli Studi di Genova, ♣ Sofia University
† Artificial Intelligence and Language Technologies Department, IICT-BAS
chiara.alzetta@edu.unige.it, {petya, kivs}@bultreebank.org,
{felice.dellorletta, simonetta.montemagni, giulia.venturi}@ilc.cnr.it

Abstract

The paper illustrates a case study aimed at identifying cross-lingual quantitative trends in the distribution of dependency relations in treebanks for typologically different languages. Preliminary results show interesting differences rooted either in language-specific peculiarities or cross-lingual annotation inconsistencies, with a potential impact on different application scenarios. [1]

1 Introduction and Motivation

The identification of cross-lingual quantitative trends in the distribution of dependency relations in "gold" treebanks is increasingly attracting the interest of the computational linguistics community for different purposes, as testified e.g. by a recently published miscellaneous book on the quantitative analysis of dependency structures (Jiang and Liu, 2018) or pilot initiatives such as the first edition of the workshop "Quantitative Syntax 2019"[2]. Among possible applications, it is worth mentioning studies aimed at acquiring typological evidence to be integrated in multilingual NLP algorithms (see Ponti et al. (2018) for a survey and the workshop "Typology for Polyglot NLP"[3]), or at detecting annotation inconsistencies to improve the quality of treebanks (see (Dickinson, 2015; de Marneffe et al., 2017) to mention only a few). While the latter is a well-established research topic, although with still many open issues, automatically acquiring typological information is still at its beginning, so automatic strategies to extract such information from corpora are needed (Cotterell and Eisner, 2017; Bjerva and Augenstein, 2018).

Multilingual resources such as the dependency treebanks developed within the Universal Dependencies (UD) project[4], thanks to the cross-linguistically consistent syntactic annotation (Nivre, 2015), fostered the development of automatic strategies to extract cross-lingual similarities and differences in shared constructions from corpora (Murawaki, 2017; Bjerva et al., 2019). Within this line of research, the paper describes a methodology for comparing treebanks of typologically different languages with the final aim of detecting and quantifying similarities and differences in multilingual treebanks analyzed from a twofold perspective: language–specific peculiarities vs cross–lingual annotation inconsistencies. To this end, we used LISCA (*LInguiStically–driven Selection of Correct Arcs*) (Dell'Orletta et al., 2013), an algorithm which has been successfully applied in different scenarios, against both the output of dependency parsers and gold treebanks. In the first case, the score returned by LISCA was meant to identify unreliable automatically produced dependency relations (Dell'Orletta et al., 2013). When used against gold annotations, LISCA was used to detect shades of syntactic markedness of syntactic constructions in manually annotated corpora from a monolingual perspective (Tusa et al., 2016), or to acquire quantitative typological evidence from a multilingual perspective (Alzetta et al., 2018b). Last but not least, it was also exploited to identify anomalous annotations (going from annotation inconsistencies to errors) from a monolingual perspective in gold treebanks (Alzetta et al., 2018a).

The methodology exploited for the present work (described in Section 2) was tested in a case study carried out on four Indo-European languages belonging to three different genera (according

[2] https://www.aclweb.org/anthology/W19-79.pdf

[3] https://typology-and-nlp.github.io/

[4] https://universaldependencies.org/

to WALS classification, Dryer and Haspelmath (2013)): Bulgarian (Slavic, BUL), English (Germanic, ENG), Italian and Spanish (Romance, ITA and SPA). UD treebanks constitute an ideal test bed for our analysis since, sharing the same annotation scheme, allow the investigation of cross-lingual similarities and differences in shared constructions. Besides similarities connected with the UD annotation strategy aimed at maximising parallelism across languages, results in Section 4 reflect shared possibly "universal" features of languages. Differences, in turn, can either reflect typologically relevant language peculiarities or highlight inconsistencies in the application of the shared annotation scheme. The paper focuses on both aspects. Section 5 concludes the paper discussing our findings and future directions of research.

Contribution. The present contribution has two main goals: we aim to show how the methodology can be used 1) to acquire quantitative evidence of cross-linguistically shared properties, and 2) to highlight divergences due either to language idiosyncrasies or annotation inconsistencies across treebanks.

2 Method

As shown in Figure 1, our methodology for exploring multilingual treebanks is articulated in the following two steps.

I) LISCA Analysis. The LISCA algorithm operates in two steps: 1) it collects statistics about a set of linguistically motivated features extracted from an automatically dependency parsed corpus (referred to as *Reference Corpus*) to build a statistical model (SM) of the language; 2) it uses the obtained SM to assign a score to each dependency relation (DR) instance, defined as a triple d(ependent), h(ead), t(ype) of dependency linking d to h, in a *Target Corpus*. Borrowing a metaphor from Jakobson (1973), we can look at the SM as encoding the DNA of the language being analysed. Note, in fact, that the features considered by the LISCA algorithm to build the SM cover, for each DR instance, a wide variety of factors, both local and global. *Local features* include e.g. the distance in terms of tokens between d and h, the associative strength linking the grammatical categories involved in the relation (i.e. POS_d and POS_h), the POS of the head governor, the type of dependency connecting d to h, and the relative linear order of d and h in the sentence. *Global features*, instead, are aimed at locating each DR within the overall sentence structure, and include e.g. the distance of d from the root of the dependency tree or from the closest or most distant leaf node, and the number of "brother" and "children" nodes of d, occurring respectively to its right or left in the linear sequence of words of the sentence. In this case study, LISCA has been used in its delexicalized version in order to abstract away from variations resulting from lexical effects, thus guaranteeing cross-lingual comparability of results. The output of LISCA consists of the list of all DRs in the Target Corpus ranked by decreasing score.

The LISCA score is a context-sensitive and frequency-based measure reflecting the degree of similarity of the "linguistic environments" in which a given DR occurs in the Reference and Target corpora: it encodes the probability to observe a DR instance occurring in a specific context on the basis of the Statistical Model constructed starting from the Reference Corpus. In more abstract terms, the LISCA score can be seen as reflecting the prototypicality degree of a specific linguistic structure: whereas higher LISCA scores identify DR instances appearing in "typical" (more frequent and likely) contexts with respect to the statistics acquired from the Reference Corpus, lower scores identify less common or even atypical DR instances of the Target Corpus. From a multilingual perspective, the comparison of the ranked DRs lists obtained from corpora of different languages can shed light on similarities and differences at linguistic and/or annotation levels. To carry out this comparative analysis, in this study the ranked list of DRs has been split into 20 intervals of equal size, henceforth "bins" (plus a further bin for the remaining ones): the first bins contain DRs presenting a high LISCA score and, conversely, the last bins contain DRs associated with low LISCA scores.

II) Ranking Exploration. We exploited CLaRK system (Simov et al., 2004) to identify and compare quantitative trends from LISCA rankings. CLaRK system work–flow is the following: firstly, each Target Corpus is converted from the CoNLL-U format[5] into XML format, then the XPath language is used to select the nodes (sentences or tokens) with the required properties. In this way we

[5] http://universaldependencies.org/format.html

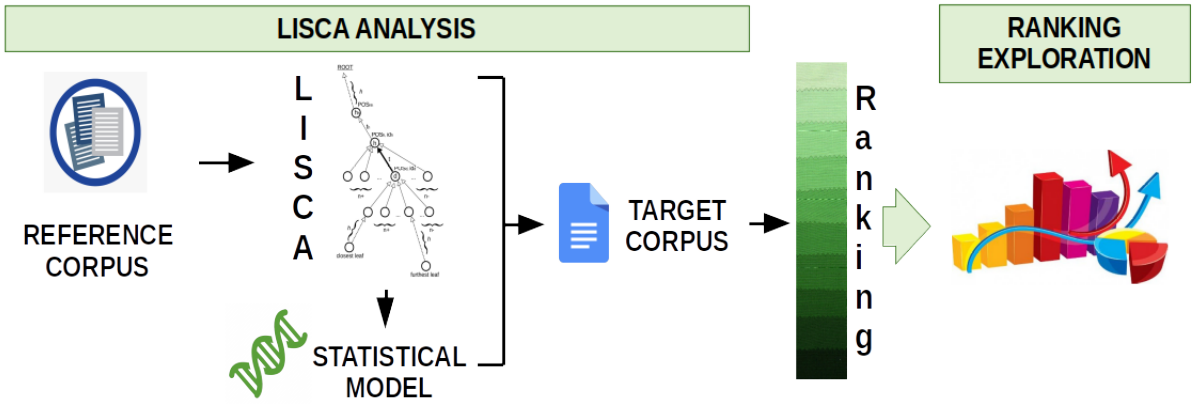

Figure 1: Method work–flow.

can define different configurations and check the distribution of the node characteristics along the DR rankings.

3 Data

For each language taken into account, two linguistically annotated corpora have been used: a large Reference Corpus and a Target Corpus.

Each *Reference Corpus* consists of a monolingual corpus of texts from the news and Wikipedia domains of around 40 million tokens, constituting a set of examples large enough to reflect the actual distribution of phenomena in the specific language. Reference corpora were morpho-syntactically annotated and dependency parsed by the UDPipe pipeline (Straka et al., 2016) trained on the Universal Dependency treebanks, version 2.2 (Nivre et al., 2017).

Target corpora correspond here to manually validated ("gold") Universal Dependencies treebanks (v2.2). Specifically, we considered the following UD treebanks:
i) English Web Treebank (254,830 tokens and 16,622 sentences) (Silveira et al., 2014);
ii) Italian Stanford Dependency Treebank (278,429 tokens and 14,167 sentences) (Bosco et al., 2013);
iii) Spanish UD treebank (547,680 tokens and 17,680 sentences) (Alonso and Zeman, 2016);
iv) UD_Bulgarian-BTB (156,149 tokens and 11,138 sentences) (Simov et al., 2005).

4 Results

Results are analysed from a twofold perspective, focusing on the distribution across the bins of different DR types and structures.

4.1 Ranking of Dependencies

As pointed out above, higher LISCA scores are assigned to DRs that show a linguistic context highly typical for the language, whereas low scores are associated with atypical (or simply less typical) syntactic structures; (un)typicality is assessed here with respect to the statistics acquired from the Reference Corpus.

As a first step of our comparative analysis, for each language we focused on the distribution of individual DRs across the 20 LISCA bins. Figure 2 reports the median bin of occurrence for all 29 shared DRs in the ranking of each language. The median bin was selected by sorting all instances of a given DR on the basis of the associated LISCA score and by identifying the median element of the ranked list: its bin of occurrence was taken as representative of the relation. Top and bottom relations (respectively at the extreme left and right in Fig.2 graph) in language-specific rankings show interesting similarities: if on the one hand DRs involving function words (e.g. `case`, `det`, `aux(:pass)`) are associated with higher LISCA scores for all languages, on the other hand special or "loose" DRs such as `orphan` and `parataxis` or clausal subjects and adverbial clauses (`csubj(:pass)`, `advcl`) all occur in the last bins, representing relations with more variable contexts across all languages. Another cross-language parallelism concerns the relative rankings of subsets of DRs: clausal complements with obligatory control (`xcomp`) are assigned a higher score with respect to the wider class of clausal complements without it (`ccomp`); the direct object relation (`obj`) precedes in the ranking the oblique argument/modifier (`obl`); and the nominal subject (`nsubj`) always precedes its

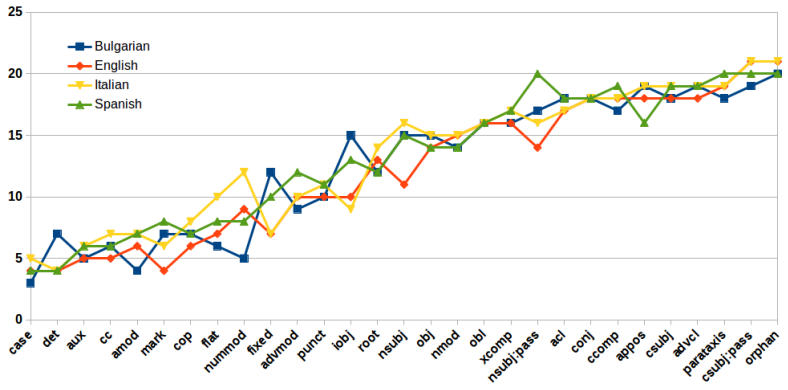

Figure 2: Median occurrence in LISCA bins of shared DRs across languages.

	Bins 1 to 10				Bins 11 to 20			
DR length	**BUL**	**ENG**	**ITA**	**SPA**	**BUL**	**ENG**	**ITA**	**SPA**
1	66.42	55.80	65.79	69.16	38.96	37.62	36.97	37.34
2	20.26	23.31	24.93	21.62	16.69	17.93	17.05	17.47
3-4	6.45	12.24	5.16	5.21	19.45	21.11	15.85	16.31
5-10	4.75	4.72	2.71	2.01	14.34	14.09	13.05	13.34
$\geq$10	2.39	3.93	1.41	1.98	10.56	9.25	17.08	15.54
# DRs	64,885	110,184	141,389	248,794	22,510	42,475	44,310	96,781

Table 1: Percentage distribution by length of DRs involving leaves in the first and in last 10 LISCA bins. For each group of bins the number of all DRs involving leaves is given.

clausal counterpart (`csubj`). It is interesting to report that the frequency of a DR seems to plays a minor role in determining the position of a given DR in the LISCA ranking: consider, for instance, the `punct` relation which is a highly frequent DR (covering around 11% of DRs in all four languages), but nevertheless it was placed in the middle part of the ranking for all languages. Looked at from this perspective, the LISCA ranking of relations - which is heavily influenced by the principles underlying the UD annotation schema - seems to reflect the parsing complexity of relations (Alzetta et al., 2020), where more complex to parse DRs are characterised by a higher variability in their contexts of occurrence.

Some interesting differences can also be reported, originating either in a) language-specific peculiarities or b) possibly inconsistent annotations across languages. Concerning a), ENG nominal subjects (`nsubj`, `nsubj:pass`) are ranked significantly higher with respect to the other three languages, all sharing the pro-drop and free word order properties; or determiners (`det`) show the same distribution for SPA, ENG and ITA in contrast to BUL, where the definite article is postpositioned and expressed morphologically, with the exception of some pronouns functioning as determiners, e.g. demonstratives. Here are two examples for Bulgarian where the first one shows the usage of the morphologically expressed postpositioned definite article (thus no explicit (`det`) relation) while the second shows the usage of a demonstrative pronoun (marked with (`det`) relation)): (1) ('Жената влезе в стаята') *(lit. Woman-the entered room-the*) and (2) ('Тази жена влезе в стаята') *(lit. This woman entered room-the*). The frequency of the examples type (1) in the treebank is about 10 times bigger than the frequency of the examples of type (2). Thus, the `nsubj` nodes modified by explicit determiner word is a rare case in Bulgarian treebank.

With respect to b), there are interesting examples, even among core UD DRs: this is the case of indirect objects (`iobj`), whose annotation criteria highly diverge across languages. The sources of dissimilarities might come partially from the annotation specifications per language about what a second argument (`iobj`) vs an adjunct (`obl`) is. If a closer look is taken into the data, it turns out that in ITA and ENG the `iobj` is typically expressed by a PRON(oun), as in these two examples: ITA: '*ti* (PRON) *ho dato*' (lit. 'I gave you'); ENG: '*causing us* (PRON) *truble*'. In ITA this represents 100% of the cases, while in ENG 84%,

whereas in SPA and BUL this relation is expressed by a pronoun in only 46.7% and 19% of the cases respectively. In Spanish, for example, the `iobj` relation is used also for NOUNs: in the Spanish example '*Obligaron al Gobierno* (NOUN) *a comprar creditos*' (lit. *Forced the Government to buy credits*) the noun is annotated as indirect object of *obligaron*, whereas in Italian the construction '*Non ho dato soldi al presidente* (NOUN)' (lit. *I didn't give money to the president*) the noun is marked as `obl` relation. In Bulgarian the `iobj` relation is used not only for marking the dative pronouns, but also for marking head NOUNs in PPs. The prevalence of this relation on NOUNs is due to the following factors: (1) the existence of long dative counterparts to short dative pronouns that consist of a preposition and a noun ('Майката даде играчка на детето') (prep NOUN) (lit. *Mother-the gave toy to child-the-DAT*); and (2) the marking of indirect complements as indirect objects, while the `obl` relation has been reserved for adjuncts ('Те продължават да участват в лотарията') (non-dative prep NOUN) (lit. *They continue to participate in lottery-the*). This suggests that different annotation criteria guide the assignment of the `iobj` DR, possibly not all of them originating in peculiarities of the language.

Other interesting examples concern the annotation of multi-word expressions and proper names (`fixed` and `flat`), which are treated differently across languages. For example, in BUL all grammatically fixed multi-words, such as complex prepositions (like с оглед на 'with regard to') or conjunctions (like за да 'in order to'), are treated as `fixed` while in Italian the annotation reflects the underlying syntactic structure, as in the case of, e.g., '*a base di*' (lit. *made of*) and '*in relazione a*' (lit. *in relation to*).

4.2 Distribution of Leaves

For each language, we investigated the distribution of DRs across the LISCA bins focusing on DRs involving leaves as dependants (henceforth *leaves*), as opposed to DRs without leave nodes (henceforth *non-leaves*). Results of this analysis are reported in Table 1. Despite minor differences, all languages share a similar trend: leaves are mostly ranked in the first 10 bins representing for Bulgarian 91.52% of the DRs occurring in them, 95.56% for English, 98,27% for Italian and 91.76% for Spanish. Interestingly, the first 6, 6, 8 and 4 bins respectively for Bulgarian, English, Italian and Spanish contain exclusively leaves. In other words, leaves are typically associated with higher LISCA scores: due to their smaller context, they are characterised by higher processing reliability. This is in line with the fact that DRs involving functional words, e.g. `case`, `det`, `aux`, etc. typically occur in the first bins (see Figure 2). On the contrary, the last 10 bins of all languages mostly contain DRs not involving leaves (68.28% BUL, 63.54% EN, 69.33% ITA, 64.54% SP). For what concerns the leaves in the second half of the bins, they turned out to be typically involved in particularly complex syntactic contexts, such as long distance dependencies or occurring in constructions that are not typical for that relation.

5 Conclusion

In this paper we presented method for studying the distribution of DRs in gold treebanks which was tested in a case study carried out on four languages belonging to three different genera. The cross-lingual comparison of the LISCA-based ranking of UD relations across the bins shows: on the one hand, shared (possibly universal) trends, concerning e.g. the similar distribution of dependencies involving leaves or of long distance dependencies, which are respectively concentrated at the top and at the bottom of the LISCA ranking for each language; on the other hand, recorded differences in the ranking of relations can be explained in terms of either language peculiarities (e.g. the pro-drop property of BUL-ITA-SPA vs ENG, or the surface realisation of definite determiners in BUL vs ENG-ITA-SPA) or potential inconsistencies in the application of the UD annotation scheme (see the case of the indirect object relation). Both types of results play a potentially key role in different scenarios, going from typology-driven multilingual NLP to the improvement of the cross-lingual consistency of treebanks.

Acknowledgements

Thanks to the anonymous reviewers for their helpful comments. This work was partially supported by *the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant num-

ber DO01-272/16.12.2019.

References

H. M. Alonso and D. Zeman. 2016. Universal dependencies for the ancora treebanks. *Procesamiento del Lenguaje Natural*, 57:91–98.

C. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2018a. Dangerous relations in dependency treebanks. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 201–210, Prague, Czech Republic, January.

C. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2018b. 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. Alzetta, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2020. Uncovering typological context-sensitive features. In *Proceedings of the Second Workshop on Typology for Polyglot Natural Language Processing*.

Johannes Bjerva and Isabelle Augenstein. 2018. From phonology to syntax: Unsupervised linguistic typology at different levels with language embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 907–916.

Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1529–1540, Minneapolis, Minnesota, June. 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, August.

Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1182–1192.

M.C. de Marneffe, M. Grioni, J. Kanerva, and F. Ginter. 2017. Assessing the Annotation Consistency of the Universal Dependencies Corpora. In *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2007)*, pages 108–115, Pisa, Italy, September.

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

M. Dickinson. 2015. Detection of Annotation Errors in Corpora. *Language and Linguistics Compass*, 9(3):119–138.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Roman Jakobson. 1973. *Essais de linguistique générale t. 2: rapports internes et externes du langage*. Les éditions de Minuit.

J. Jiang and H. Liu. 2018. *Quantitative Analysis of Dependency Structures.* De Gruyter Mouton, Berlin, Boston.

Yugo Murawaki. 2017. Diachrony-aware induction of binary latent representations from typological features. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461.

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*.

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, April.

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, A. Simov, H. Ganev, K. Ivanova, and I. Grigorov. 2004. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238.

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).

E. Tusa, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2016. Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessitá. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, pages 3–16, Napoli, Italy, December.