Italian L1 learner treebank

Marta Sartor, ILC-CNR (Italy)

Key words: learner production, L1 learners, Italian

Learner corpora represent a crucial resource for language acquisition studies, both within and beyond the NLP community: it is therefore not suprising that, since the nineties, they have received increasing attention (Di Nuovo *et al.*, 2022). However, as is often the case, the majority of resources focus on the English language (Barbagli *et al.*, 2016). This discrepancy is particularly evident in the availability of corpora for L1 learners, with fewer resources compared to those available for L2 learners (*ibidem*). Learner treebanks, therefore, are even more scarce: for what concerns Italian in particular, only two treebanks containing learner productions are available: VALICO-UD (Di Nuovo *et al.*, 2022), by L2 learners, and MarkIT (Paccosi and Aprosio, 2022), by L1 learners. However, while VALICO-UD contains annotated sentences from an L2 learner corpus, VALICO (Barbera *et al.*, 2004), and is thus a fully fledged learner treebank, MarkIT is actually a treebank for syntactically marked structures taken from high school students' essays, which are only coincidentally productions by L1 learners. Therefore, there is no actual L1 learner treebank available for Italian.

This proposal outlines the creation of a new treebank focusing on Italian L1 learners' corpora, aiming to address this gap. The treebank will be constructed by annotating datasets from two ongoing research projects: LUCET and Verba et Acta. LUCET (LingUistic Complexity Evaluation in educaTion) is an Italian national project (PRIN 2022) which investigates linguistic complexity and its interaction with processing difficulty for the Italian language; to this goal, the project will also involve collecting written productions from high school students, both as short texts and as singular sentences elicitated to contain a linguistic structure of particular interest. The Verba et Acta project, on the other hand, involves the Università Roma Tre, the ILC-CNR research group and the Regional School Managment of Marche. The project investigates the potential connection between writing about practical activities and their execution, monitoring if increased linguistic competence on the subject can positively impact the execution itself. To this end, short texts are repeatedly elicited from vocational schools students.

The written productions collected through these projects, both of which my research group participates in, could be made available as an UD treebank of L1 high school students written productions. The annotation would be obtained with a semi-automatic methodology, already validated in noumerous studies (e.g. Alzetta *et al.,* 2017; Alzetta *et al.*, forthcoming), consisting in leveraging an algorithm to pinpoint critical nodes in automatic annotation to manually identify and correct systematic errors. The LISCA (Linguistically-driven Selection of Correct Arcs) algorithm (Dell'Orletta et al, 2013) scores automatically generated dependency relations on the basis of the reliability of their annotation. Relations showing a lower plausibility score can then manually inspected to identify possible recurring patterns that may indicate potentially erroneous annotations requiring manual correction; the patterns can then be translated into heuristics to identify potential errors across the entire corpus, which are then manually evaluated and eventually corrected. This effectively restricts the search space for errors and helps identify patterns of systematic recurrent errors that are generated by a parser (Alzetta *et al.,* 2017).

Learner productions, being low-represented and possessing their own linguistic specificities, require careful and well-thought-out annotation processes. Deepening my understanding of annotation systems, choices, and the tools supporting treebank construction will facilitate a more accurate and efficient annotation process, ensuring the quality and consistency of annotations in the proposed treebank. Additionally, acquiring insights into annotation systems at the outset can help prevent errors early in the resource construction, avoiding costly and less accurate adjustments later on.

Furthermore, participating alongside other young researchers, especially those working on lower-resourced languages or genres, promises to be a truly enriching experience. Engaging with new perspectives will stimulate critical reflection on annotation systems, potentially leading to better-informed decisions in ambiguous or complex scenarios during the annotation process.

Interesting arguments of debate relating the project could be, for example, what to mark as error: only inacceptable linguistic forms and structures or also grammarly incorrect but widely used forms? Where should the line between the two be drawn, and how to best annotate them to maintain interlinguistic comparability as the Universal Dependencies framework requires?

The project is still, as yet, a proposal. However, data collection has already stared for Verba et Acta and will soon start for LUCET: preliminary analysis of the automated annotation will be possible shortly, allowing some insights on possible shortcomings of automatic annotators and/or unexpected variety-specific peculiarities which will require adjustments. In the meantime, the focus can be on dataset construction and annotation choices.

References:

ALZETTA, Chiara, et al. Dangerous relations in dependency treebanks. In: *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*. 2017. p. 201-210.

ALZETTA, Chiara, et al. ParlaMint-It: an 18-karat UD treebank of Italian parliamentary speeches. In: Language Resources and Evaluation. Forthcoming.

BARBAGLI, Alessia, et al. CItA: An L1 Italian learners corpus to study the development of writing competence. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 2016. p. 88-95.

BARBERA, Emanuele Ferdinando, et al. VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online): una presentazione. *ITALS*, 2004, 2: 7-18.

DELL'ORLETTA, Felice; VENTURI, Giulia; MONTEMAGNI, Simonetta. Linguistically-driven selection of correct arcs for dependency parsing. *Computación y Sistemas*, 2013, 17.2: 125-136.

DI NUOVO, Elisa, et al. VALICO-UD: Treebanking an Italian Learner Corpus in Universal Dependencies. *IJCoL. Italian Journal of Computational Linguistics*, 2022, 8.8-1.

PACCOSI, Teresa; PALMERO APROSIO, Alessio. It is MarkIT that is new: An italian treebank of marked constructions. In: *Proceedings of CLiC-it 2021 Italian Conference on Computational Linguistics*. 2022.