Training School with UniDive

Building a Treebank of Italian L1 Learner Productions
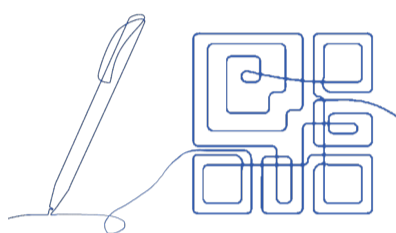
Date: 8-12 July 2024

Marta Sartor
CNR-ILC - Pisa
ItaliaNLP Lab – www.italianlp.it

Data Collection

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 elicited to contain a linguistic structure of particular interest.
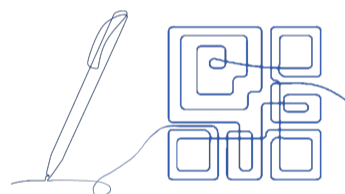
The **Verba et Acta** project, on the other hand, involves the Università Roma Tre, the ILC-CNR research group and the Regional School Managment of Marche. The project investigates the potential connection between writing about practical activities and their execution, monitoring if increased linguistic competence on the subject can positively impact the execution itself. To this end, short texts are repeatedly elicited from vocational schools students.
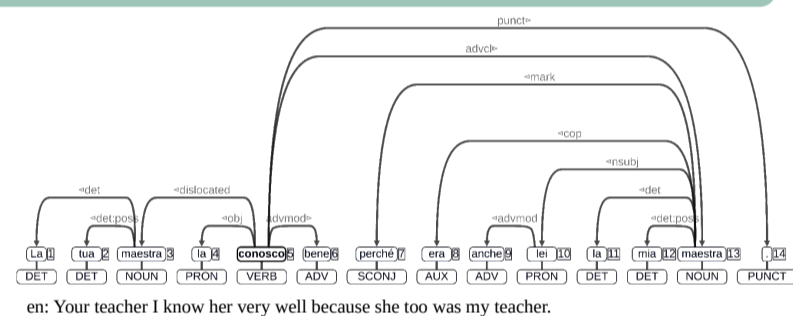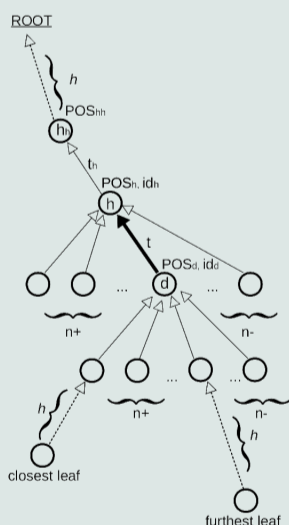
Annotation Methodology

1) Automatic Annotation

LuCET + [pen icon] → Stanza →

La [1] tua [2] maestra [3] la [4] **conosco** [5] bene [6] perché [7] era [8] anche [9] lei [10] la [11] mia [12] maestra [13] . [14]
DET DET NOUN PRON VERB ADV SCONJ AUX ADV PRON DET DET NOUN PUNCT

en: Your teacher I know her very well because she too was my teacher.

2) Semi-Automatic Revision using LISCA

RANKING →

Random error identification → Manual correction

Error pattern definition → Manual correction

3) Revision Validation

Random sentence sampling → Manual inspection and correction → Ready for release!