From manuscript to syntactic tree: the long journey of mathematical Latin

In this paper, we want to present the work carried out to produce syntactic annotation of Neolatin mathematical texts, in particular of the Latin translation of Archimedes' works by Jacopo da San Cassiano (XV century), and the first results emerging from the analysis. The goal is to illustrate a full process of digital analysis of a Neolatin text, from the manuscript to the syntactic tree, and to discuss the specific challenges linked to the automatic processing of a technical text. The pipeline entails the re-use of existing data, tools, and models combined to reach the general research objective.

Part I: text digitization and editing

In the first part of the paper, we will present the work of Jacopo da San Cassiano, who translated most of the works of Archimedes that circulated in Western Europe at the time, and the digitization process of the text. Recent studies (d'Alessandro-Napolitani 2012) have identified the Jacopo autograph with ms. Paris, Bibliothèque Nationale de France, *Nouv. acq. Lat.* 1538. The text shows up as a draft, and it has been critically edited by B. Sisana, in the frame of her Master thesis, making use of a LateX package (MauroTeX) developed specifically for the digital editing of Latin mathematical texts. Editing a Latin mathematical text in the form of a draft entails specific challenges (rendering of mathematical notation, acknowledging the different stages of hesitation) which are addressed by the package. For this pilot study, we have selected one treatise, *The spirals*, and extracted the main text from the .tex file.

Part II: lemmatization and POS tagging

The second part of the presentation will deal with the linguistic annotation process. Technical and scientific Latin has generally been overlooked in the creation of digital Latin corpora (this is now changing, see for instance Grotto et al. 2021). Nonetheless, the interest of data creation and annotation is double-fold: first, these texts open to different variants of Latin, and second they give access to the way in which key-works for the History of Science were shared and understood across time. The task proves cumbersome: no training data are available, traditional tools perform poorly and even manual annotation requires to develop *ad hoc* solutions for the specific text.

First, the LASLA-Pie model developed by Th. Clérice has been used to automatically lemmatize and morphologically tag the text (ref. to the model). The model has been adapted by Th. Clérice to tackle some specific features of the mathematical text, namely the presence of letters to indicate mathematical objects that might generate mistakes in the lemmatization/tagging as they are genre-specific features. The results for the first 3000 tokens have been manually corrected using the interface Pyrrha. The main errors in the lemmatization and POS tagging will be discussed here.

Part III: Dependency parsing

In a second stage, we trained a Latin parser using the MaChamp library (van der Groot et al. 2021), based on the model developed by Stymne et al. 2018 starting from the UUPArser (de Lhoneux et al. 2017). The goal of the model is to tackle the issue of the non-homogeneity of available UD treebanks, by creating a treebank embedding to represent the treebank to which the sentence belongs. For this specific case, a multitask model was trained, first to predict POS tagging and then to parse the text. The training set was constituted by available treebanks of a cluster of ancient languages, and ca 140 sentences from Archimedes' *Spirals*. For the POS prediction task, we used also the L.A.S.L.A. files (now available online), in their UD format (processed by Flavio Cecchini, LiLa ERC). We will discuss the performance of the model, focusing in particular on the issues related to the mathematical features of the text: different annotation strategies have a clear impact on the features of the resulting annotation. We will observe in particular whether adding training data from the Archimedean corpus significantly changes the performance. As a conclusion, we will discuss the general feasibility of described pipeline for the entire Archimedes corpus and provide some first linguistic observations that can be inferred from the parsed text.

References

Paolo d'Alessandro, Pier Daniele Napolitani (2012). *Archimede Latino: Iacopo da San Cassiano e il "Corpus" Archimedeo alla metà del Quattrocento, con edizione della "Circuli dimensio" e della "Quadratura parabolae."* Paris: Les Belles Lettres, 2012.

Francesco Grotto, Rachele Sprugnoli, Margherita Fantoli, Maria Sini, Flavio Massimiliano Cecchini, Marco Passarotti (2021). The Annotation of Liber Abbaci, a Domain-Specific Latin Resource. Proceedings of the Eighth Italian Conference on Computational Linguistics (clic-it 2021). Milan, Italy. 2021. https://doi.org/10.5281/zenodo.5773777

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, Barbara Plank. (2021). Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. 10.18653/v1/2021.eacl-demos.22

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre (2018). Parser Training with Heterogeneous Treebanks, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 619–625. Melbourne, Australia, July 15 - 20, 2018.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. (2017). From raw text to universal dependencies - look, no tags! In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pages207–217, Vancouver, Canada.