Timo KORKIAKANGAS

Late Latin Charter Treebank: Building an extensive treebank of early medieval documentary Latin

My talk presents the Late Latin Charter Treebank (LLCT), a dependency treebank of early medieval Italian formulaic documentary texts ("charters") written in non-standard Latin. The talk explains how LLCT is designed to enable complex linguistic and philological research settings with diachronic and diatopic dimensions.

The version 1 of LLCT (completed in 2014) was based on 519 documents from Tuscany of 714-869 AD. Version 2 (2016-2018) added 521 Tuscan documents, expanding the time frame up to 897 AD. Version 3 (2019-2021?) will again extend the corpus chronologically, by adding c. 100 Tuscan documents from the 10th century, but also geographically, by adding c. 160 documents from northern and southern Italy. The final number of documents will be c. 1,300 (c. 600,000 tokens). Only original documents or contemporary copies are included.

I will present the workflow from transcribing documents/scanning editions up to the completed treebank. This entails explaining the choices made during the initial selection of documents, as well as decisions related to their annotation. Each document is provided with metadata (various identifiers, date, scribe name, scribe title, writing place, geocoordinates, document type). By the completion of LLCT3, LLCT will be enriched with further metadata concerning palaeographical features of the documents.

LLCT1 was annotated manually in the Alpheios annotation environment with the help of Morpheus morphology tool. LLCT2 and LLCT3 were lemmatized with a simple Visual Basic script and morphologically and syntactically tagged with TnT tagger and Malt Parser, respectively, with LLCT1 serving as the training set. The lemmas and morphological tags were subsequently corrected manually in Excel and the syntactic tags in the Arethusa annotation framework.

The linguistic annotation mainly follows the Guidelines for the Syntactic Annotation of Latin Treebanks (v. 1.3, Bamman & al. 2007). The Guidelines are designed for Classical Latin while the non-standard Latin of LLCT requires some specifications regarding syntax and, particularly, regarding morphology. The latter are based on an etymology principle that reduces the LLCT morphology to Classical Latin grammatical apparatus (Korkiakangas & Passarotti 2011: https://jlcl.org/content/2-allissues/12-Heft2-2011/16.pdf). LLCT also contains a philological annotation layer which includes diplomatic segmentation (free vs. formulaic parts), morphologically ambiguous abbreviations, normalized form, and Levenshtein distance between each attested and normalized form.

The ambiguity of certain non-standard Latin morphemes as well as the thereon consequent etymology principle make it challenging to align LLCT with other treebanks, even those annotated with the Guidelines. An illustrative example is the treatment of OBJ, exclusively reserved to encode direct and indirect objects in LLCT. This is because 1) LLCT was originally meant to be used to study subject case variation only, 2) morphological case does not always distinguish between different complements in the Latin of LLCT, 3) complement/adjunct attribution is notoriously ambiguous with formulaic documentary language.

LLCT is in PML format. The current version of LLCT1 is available at Zenodo (doi.org/10.5281/zenodo.1197357) and that of LLCT2 upon request from author. Both will be shared at the FIN-CLARIN repository as soon as the student-corrected syntactic annotation of LLCT2 is verified and parts of the annotation of LLCT1 checked.