Giuseppe CELANO

Standoff Annotation for the Ancient Greek and Latin Dependency Treebank

This contribution presents the work in progress to convert the Ancient Greek and Latin Dependency Treebank (AGLDT) into standoff annotation using PAULA XML. The AGLDT contains morphosyntactic annotation for 557,022 AG tokens and 79,697 Latin tokens. More precisely, the morphological annotation consists in POS tagging and morphological features, while the syntactic one in labeled head-dependent relationships.

With an increasing number of annotations of any kind, it becomes more and more urgent that annotations related to the same texts be added standoff. Standoff annotation has been standardized in the Linguistic Annotation Framework (LAF) (ISO TC37 SC4), which is particularly indebted to the Corpus Encoding Standard (CES), where standoff annotation was referred to as "remote markup".

Standoff Annotation consists in adding any kind of annotation in separate documents, which are ultimately linked to a main text, the so-called "base text", which is meant to be unchangeable. References occur via a graph-based system of IDs, which allows an annotation layer (contained in a separate file) to be linked to another annotation layer (contained in another separate file). All the annotations/files create a labeled directed acyclic graph, whose root is represented by the base text. Contrary to in-line annotation, where more kinds of annotation are taken together within the same file, standoff annotation enables easy interoperability and extension, in that single annotation layers can reference other layers of annotation independently, thus overcoming the problem of conflicting hierarchies. Moreover, standoff annotation also allows addition of different annotations of the same kind to the same text (e.g., two different interpretations of the POS tag for a given token).

An example of the advantage of standoff annotation for the Ancient Greek and Latin texts is provided by different segmentations a given text requires for syntactic and prosodic analyses. For example, the sentence, the largest unit for syntactic analysis, may or may not coincide with a verse, which is a prosodic unit. Mismatches between syntactic and prosodic units ask for different segmentations and therefore for a system where all annotation layers are kept distinct, but each of them can be linked to another annotation layer, with all layers ultimately referencing an unchangeable base text.

Standoff annotation for the AGLDT is achieved using PAULA XML, which is an open access format following the LAF principles. PAULA XML has already been successfully employed in a number of annotation projects for historical languages (http://www.laudatio-repository.org/repository/). In the present contribution I show the pilot case of Caesar's *De Bello Civili*: for the first time, the base text is not a TXT file, where regions are referenced via character offsets, but a TEI XML file, where regions are identified via XPath/XPointer expressions. I detail the PAULA XML files created for tokenization and sentence detection, which are necessary to add morphosyntactic annotation.