Freitag
10.03.2017
11:30 – 12:00
B3 1, 0.12

TEITOK: Combining language and linguistic information without compromise

Maarten Janssen
CELGA-ILTEC, Universidade de Coimbra
maarten.janssen@campus.ul.pt

AG4

TEITOK is an online XML-based corpus framework using a strategy to combine the needs of various users of (historic) corpora, by allowing multiple orthographic realizations of the same text, such as a paleographic and a regularized form to serve both philological and computational needs. It uses an inheritance tree for forms, so that only distinct forms have to be stored explicitly. The different forms are modelled as features over XML nodes (tokens). This facilitates not merely having two different forms, but as many as required, allowing to also include a diplomatic form, a transliterated form, etc.

In such multi-layered texts, not only the orthography changes, but also the number of tokens: normalisation can both merge and split words. And for linguistic processing, it is often required to split contractions into multiple words, or groups multiple words into a MWE. To solve this, TEITOK first of all allows spaces inside tokens at any level. And secondly, it allows a single token to contain more than one grammatical word inside, where the linguistic tags are modelled over the grammatical words. In this fashion, the "word" *doutra* can be normalized to *de outra*, and then provided with two grammatical words, the first a preposition, and the second a pronoun.

TEITOK adds the token nodes inline in TEI documents that can contain rich typographic annotations, notes, facsimile images, etc. And TEITOK allows adding standoff annotation files for complex annotations. The framework contains a collection of GUI tools to manage the resulting XML files, which become so heavily annotated that editing the raw files becomes unfeasible. It also contains a POS tagger that tags directly over XML files, and can be used with existing parameters, or easily trained on the corpus itself, and in the interface all annotations can easily be corrected. It also provides a searchable version of the corpus using Corpus Workbench. The resulting corpora have proven to be of use not only for corpus/computational linguists, but also for a range of other researchers.