An integrated tool for annotating historical corpora

Pablo Faria[1] (Unicamp), Fábio Kepler[2] (USP) & Maria Clara Paixão de Sousa (USP)
pablofaria@gmail.com fabio.kepler@gmail.com mariaclara.ps@gmail.com

Brief description

E-Dictor is a tool which aims to embody the whole process of electronic encoding of ancient texts, which includes its transcription, the application of *levels of editions*, and assignment and revision of part-of-speech tags. It works also as a WYSIWYG interface to encode text in XML format. Preliminary results show a decrease of at least 50% on the overall time taken by the manual editing process.

Background

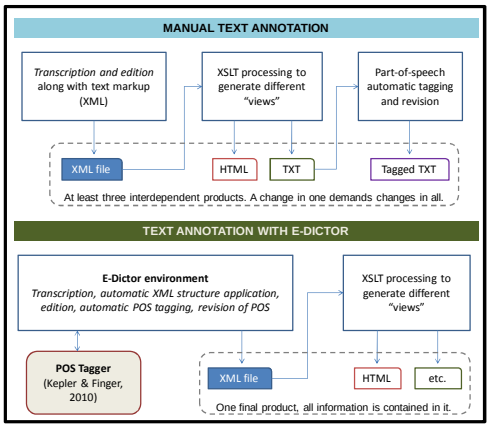
The modernization of spellings and standardization of graphematic aspects, during the first years of Tycho Brahe Parsed Corpus of Historical Portuguese (CTB) (Cor, 2010), made texts suitable for automated processing, but caused the loss of important features from the original text for the historical study of language. This tension has led to the project "Memories of the Text" (Paixão de Sousa, 2004), which sought to restructure the Corpus, based on the development of XML annotations, and to take advantage of the core features of this type of encoding, for example, XSLT processing. The annotation system was applied to 48 Portuguese texts (2+ million words), which allowed keeping philological informations while making the texts capable of being computationally treated in large-scale. Since 2006, the system has been applied by other research groups, notably the Program for the History of Portuguese Language (PROHPOR-UFBA). The system, then, met its initial objectives, but there were issues with respect to reliability and ease of use.

The manual text markup in XML was challenging to some and laborious for everyone. The basic edition process was: transcription in a text editor, application of the XML markup (tags plus philological edition), generation (from this XML file) of a standardized plain text file to submit to automatic part-of-speech tagging, and revision of both files (XML and tagged). All in this process, except for text tagging, was manual and thus subject to failures, demanding constant and extensive revision. The need for an alternative, to make the task more friendly, reliable, and productive, became clear. In short, two things were needed: a friendly interface (WYSIWYG), to prevent the user from dealing with XML code, and a way to tighten the whole process (transcription, encode/edition, POS tagging and revision).

Key features

Besides some common options (e.g., Save As, Search & Replace, and others), **E-Dictor** provides:

- Automatic XML structure generation;
- Part-of-speech automatic tagging (Kepler & Finger, 2010);
- Exporting routine of the encoded text, and;
- Exporting routine of the lexicon of editions (HTML and TXT/CSV).



UNDERSTANDING "LEVELS OF EDITION"

The original spelling and graphematic characteristics of older texts, hinder the subsequent automatic processing. Thus, it is needed a degree of interference higher than that acceptable for a (philological) semi-diplomatic edition. E-Dictor, in order to keep the original characteristics and to include these interventions, provides *levels of editions*, which work like layers.

[original]
REINANDO aquelle muy catho- lico & ſereniſſimo Principe elRey Dom MANVEL , fezẽ hũa frota pera a India de que hia por capitam mór Pedralua- rez Cabral

[resegmented]
REINANDO aquelle muy catholico & ſereniſſimo Principe el-Rey Dom MANVEL , fez-ſe hũa frota pera a India de que hia por capitam mór Pedr aluarez Cabral

[graphematic]
REINANDO aquelle muy catholico & serenissimo Principe el-Rey Dom MANUEL , fez-se hũa frota pera a India de que hia por capitam mór Pedr alvarez Cabral

[modernization]
Reinando aquele mui católico e sereníssimo Príncipe el-Rei Dom Manuel, fez-se uma frota para a Índia de que ia por capitão mór Pedro Álvares Cabral

Encoding flexibility

A key goal of **E-Dictor** is to be flexible enough so as to be useful in other contexts of corpora building. To achieve this, the user can customize it to suit his needs. The most prominent options are: *levels of edition* for tokens; *subtypes* for 'section', 'paragraph', 'sentence', and 'token' elements; and the list of POS tags to be used in the morphological analysis. Finally, in the 'Metadata' tab, the user can create suitable metadata fields as needed by each project.

Discussion

The dificulties of encoding ancient texts in XML, using common text editors, had shown that a tool was necessary to make the process efficient and friendly. This led to the development of **E-Dictor**, which, since its earlier usage, has shown promising results. Now, the user does not even have to know that the underlying encoding is XML. His concern turns to be only with philological and linguistics aspects.

E-Dictor led to a decrease of about 50% in the time required in the process. The improvement may be even higher if we consider the revision time. One of the factors for this improvement is the *better legibility* the tool provides. The XML code is hidden, allowing one to practically read the text without any encoding. To illustrate the opposite, the screenshots bellow show the common edition interface", before **E-Dictor**. Note that the content being edited is just "Ex.mo Sr. Duque".

Finally, the *integration of the whole process* into one and only environment is a second factor for the overall improvement, for it allows the user to move freely and quickly between "representations" and to access external tools transparently.

Other available tools

A survey in the internet led to some interesting tools which did not fit our needs but worth mentioning:

- *Multext* (may be outdated), at http://aune.lpl.univ-aix.fr/projects/multext/.
- *CLaRK*, at http://www.bultreebank.org/clark.
- *Xopus* (WYSIWYG), at http://xopus.com/.
- *oXygen* (WYSIWYG), at http://www.oxygenxml.com/.

REFERENCES

[Cor2010] IEL-UNICAMP and IME-USP, 2010. Corpus Histórico do Português Anotado Tycho Brahe, http://www.tycho.iel.unicamp.br/

[Kepler&Finger2010] F. N. Kepler & M. Finger. 2010. *Variable-Length Markov Models and Ambiguous Words in Portuguese*. Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas, p.15-23.

[Paixão de Sousa2004] Maria Clara Paixão de Sousa. 2004. *Memórias do texto: Aspectos tecnológicos na construção de um corpus histórico do português*. Projeto de pós-doutorado – FAPESP Unicamp.

[1] Thanks to FAPESP, n. 2008/04312-9, for funding part of the development of E-Dictor.

[2] Thanks to CAPES for the scholarship granted during the initial part of this work.