CARDS: A Research Tool for Linguistics and History

Ana Guilherme, Leonor Tavares, Mariana Gomes
CLUL, Centro de Linguística da Universidade de Lisboa
Avenida Professor Gama Pinto, 2, 1649-003 Lisboa – Portugal
Telefone: +351 21 790 47 00 – Fax: + 351 21 796 56 22
Courriel: cardsclul@gmail.com

ABSTRACT

Le Projet CARDS (Cartas Desconhecidas - Lettres Inconnues) est un outil pour des recherches historiques et linguistiques qui rendra accessible un archive on-line de lettres portugaises, écrites entre la fin du 16ème siècle et le début du 20ème.

L'objectif principal du projet CARDS est de montrer l'importance des lettres privés comme des objets historiques et de promouvoir leur étude à partir d'une perspective linguistique. Voici une façon de contribuer pour récupérer, réutiliser et rendre vif une partie éclatante de l'héritage culturel portugais. De façon à rendre ce corpus historique plus accessible et plus facile à rechercher, il a été conçu en guise d'une édition d'une base de donnés on-line.

1. THE PROJECT CARDS

CARDS (Cartas Desconhecidas – Unknown Letters) is a tool for linguistic and historical research that will make available an on-line edition of Portuguese letters, written between the end of the 16th century and the beginning of the 20th. In the present stage, 1300 manuscripts have already been transcribed and many more have been identified. The source documents are private original letters written by speakers from all social backgrounds (popular ones included) and the main covered topics are family, love, friendship, extortion and political business. These letters were kept within the judicial files of "Casa da Suplicação" (the Lisbon Appeal Court) and "Tribunal do Santo Ofício" (the Portuguese Inquisition) because of their material importance for condemning or acquitting accused people. Both Courts have their collection stored at the National "Torre do Tombo" Archives, in Lisbon.

These documents constitute a relevant proof of the past day by day communication. They can open the door to many study themes such as: social strata, forms of inter-personal relations, social or psychological conditions, language formalities, the use of rhetorical conventions in language uses, the non-standard Portuguese varieties in history, oral language, and so on. There are many possibilities that can be approached. We must bear in mind that when there were no telephones, the letters were the best way to send a message. We can even study them by comparison with the text messages from a mobile phone.

The main objective of the project is to bring the attention to the importance of private letters as historical documents and to promote their study so that these documents, traditionally considered of minor importance, can be brought to life and honoured as a source of knowledge.

2. EDITION METHOD AND TOOLS

2.1. XML format and TEI-DALF mark-up system

The project started the letter's edition by trying a Microsoft Office Access database format. Since several technical problems came up by the end of a trial period, we decided to switch to the XML language and the TEI mark-up conventions.

XML is a mark-up language which is widely used in the World Wide Web because of all the processing and encoding advantages it offers. It generates hierarchically organised text files which are machine-readable and human-readable, and thus can serve several purposes. XML is often compared with HTML but it has some significant differences, such as:
(i) it is extensible, so the set of tags is not limited and it can be defined by the user;
(ii) it must obbey to the definitions of two types of files containing what we could call a "deep grammar", and a "surface grammar"; these are the DTD, the Document Type Definition, which establishes hierarchy relations, element-tags and attribute-tags, and the XLT, the *stylesheet*, which transforms the tagged file into an edited text;
(iii) lastly, XML files are readable, without information loss, by all types of software, which makes them easy to be convertible into formats different from the original one.

TEI stands for Text Encoding Initiative and it is a consortium that develops guidelines that specify encoding methods for texts to be exchanged mainly in the Human Sciences academic field. It is a matured means of electronic textual edition and it is widely accepted when it comes to philological data in Human Sciences environments. The TEI conformat project that served as an example for CARDS was DALF (Digital Archive of Letters in Flanders) since it is specifically concerned with the edition of epistolary data. The DALF project's DTD is in fact the one we use in the CARDS project.

2.2. CARDS' method

In our database we have three types of files and the method that we are currently following goes like this:

(i) Each letter manuscript receives an electronic textual edition within an independant XML file. For each file a code number is assigned (e.g. CARDS0003 refers to the third letter transcribed).
(ii) An image file with the letter's facsimile (digitalized in the JPEG format) is linked to the electronic edition by means of an anchoring label within the XML file.
(iii) Two biographic registers are entered in a separate XML database (CDD.xml, i.e., CARDS Demographic Database) with the letters' author and addressee social profiles (name, birth date, occupation and social status, etc) also anchored inside the XML edition.

2.3. An XML TEI-DALF mark-up document

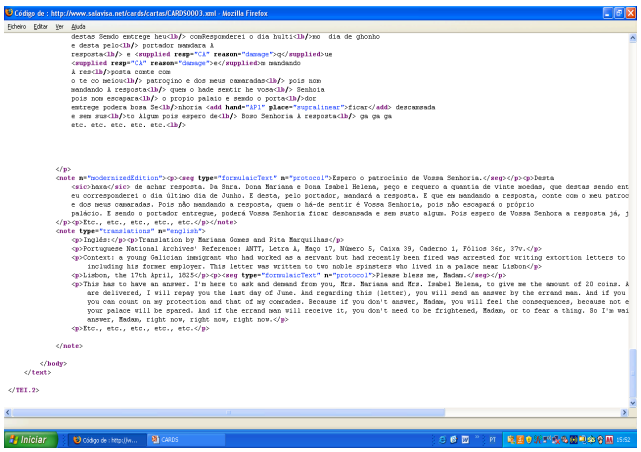
Each file starts with important declarations referring its electronic nature and the grammar it uses. Then follows the root element, the one that presides to the the TEI conformant hierarchy: <TEI.2>...</TEI.2>.

The header <teiHeader> is one of the main important parts of the file and contains all the extra-textual information about the letter's discourse, with the appropriate adjustments that DALF introduced to the TEI mark-up system. These elements refer to the people and the institutions involved in the editorial project, the transcription rules we follow, a brief commentary on the letter from the episodical, the historical and the linguistic perspectives, details on the extralinguistic context of the letter's text and on the physical characterisation of the letter's document.

The element text, the second main part of a XML file, has the semi-palaeographic edition of the letter made available by project CARDS, and, in some cases, a modernized edition also.

2.3.1. The edition method Concerning the edition methodology, first, the investigator transcribes the textual elements of the superscript (address and, sometimes, salutation). Then follows the letter's body which is divided into three parts:
(i) Opener: it has the opening elements, such as address, date and salutation;
(ii) The body of the letter;
(iii) Closer: it contains the closing elements, such as address, date, salutation and signature.

The main important elements that the transcribers mark up in the body are the formulaic text, authorial emendations and deletions, lacunae caused by external agents, hard decipherings and editor's conjectures.



Picture 1: The text's body in xml

2.4. CDD – CARDS Demographic Database

In this database we register information with biographic data concerning the letters' participants, for example: name, occupation, social status, physical description, marital status, the date when the participants received and sent letters and the episodes of their lives we could recover from the archives' documentation. The project CARDS hopes this database will prove to be very useful when we try to use the whole letter's edition as a corpus to study language change in the history of Portuguese.

3. CARDS' WEBSITE

The main goal of our project is to make available on-line thousands of private letters until the end of 2009. This achievement is possible by the construction of a website that can be accessed by everyone (as the lay public or the scientific community). With this website we have a triple objective: interfeering in the linguistic knowledge (focusing the study of language change), philological knowledge (focusing on the reflection of the differences between electronic and traditional editions) and in the historical knowledge (focusing on a modality of cultural history). CARDS' website will be attached to the Linguistics Centre of the University of Lisbon's.

We defined as imperative a presentation of a global introduction to the website, a list of the letters (showing its participants, date and a summary of each). Then, one of the most significant features the website has is the search tool: it guarantees a wide open range of linguistic and historical data that can be traced from this tool.

3.1. The main parts of CARDS' website

3.1.1. The letters The identification of each letter is maintained from the xml file, which is the number of letters edited: each transcriber has a range of 1000 numbers to inpute to the letters transcribed, so that there won't be any repetitions, because a unique code is assigned to each singular letter. The first 1000 numbers will have the Microsoft Office Access files ascribed in the first years of transcription before the project began.

There is a title area, which has an inventory of the charateristics of each letter, physically, and situates it in the whole CARDS archive – identifying parameters covering documental, linguistic and historical searching interests, as well as the original's facsimile, when possible. The letter participants' section (the ones who wrote or received letters) has a small biography of the intervenient and the list of letters in which the author or receiver takes place. This feature makes it possible to have a complete communication between writer/receiver.

Picture 2: A letter on CARDS' website

3.1.2. The search tool This is one of the most important feature displayed on our website because it ensures a free data search, whichever the area its user belongs to.

One can search for: keywords in the proper text of the letter (that have been given by specialists in linguistics, culture, and history for each letter particularly), year (of the letters), and by the name of the author or receiver of any letter. The search tool puts in practice the overall objective of CARDS project: that a user can search for a language characteristic through historical time.

4. CONCLUSION

Once it is ready for public presentation, CARDS will be a fundamental research tool for historical and linguistic purposes. This includes the cataloguing, philological transcription and facsimile reproduction of the letters, in order to offer general and professional publics an easy access to their form and contents.

It will allow for a wide study of Portuguese language variation and change across the entire Modern Ages and the entire social *spectrum*. The theory on grammar competition, on sociolinguistic variation and on the communicative efforts made by different social subjects will collect invaluable information from the processing of these non-standard language sources. Besides, it will be a support instrument for social-historical researchers, not used to such richness of private information on lives of past societies' members.

REFERENCES

[TEI07] Text Encoding Initiative, TEI:P5 Guidelines (2007)

http://www.tei-c.org/Guidelines/P5/index.xml (viewed on August 2008)

[Van04] Vanhoutte, Edward and Branden, Ron Van den, (2004), "Presentational and Representational Issues in Correspondence Reconstruction and Sorting" Mats Dahlström, Espen S.Ore, & Edward Vanhoutte (eds.), *Electronic Scholarly Editing-Some Northern European Approaches*. A Special Issue of *Literary and Linguistic Computing*, 19/1, pp.45-54.

[Van02] Vanhoutte, Edward & Branden, Van den, (2002), *Dalf guidelines for the description and encoding of mofern manuscript material*, Gent: CTB.