Projects CARDS and FLY: two multidisciplinary projects within Linguistics

Mariana Gomes, Ana Rita Guilherme, Leonor Tavares, Rita Marquilhas
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2, 1649-003 LISBOA
E-mail: mariana.gomes@clul.ul.pt, arbg@clul.ul.pt, lactavares@yahoo.com.br, rmarquilhas@fl.ul.pt

Abstract

This paper concerns the presentation of two projects that aim to make available an online archive of 4,000 original private letters, mainly having in mind research in Linguistics (Corpus Linguistics, Historical Linguistics, Pragmatics, Sociolinguistics, General Linguistics), History and Sociology. Our corpus is prepared for each research area and provides a diachronic archive of the Portuguese language. Projects CARDS and FLY have the main goal of making available an online electronic edition of each letter, which is completely open source, searchable and available. Users can search for an individual letter, a text by type, a group of letters by year or even the whole archive as a corpus for research or other purposes. The means of corpus presentation is a multimodal framework, since it joins together both the manuscript's image and the written text: the letter's material representation in facsimile and the letter's digital transcription. This editing method allows for the possibility of creating an annotated corpus where the textual unity is not lost.

Keywords: Digital Humanities, private letters, electronic edition.

1. What we do

Projects CARDS (Cartas Desconhecidas – Unknown Letters) and FLY (Forgotten Letters, Years 1900-1974) are two multidisciplinary projects carried out by the Linguistics Centre of the University of Lisbon. Since 2006, we have collected and transcribed private Portuguese letters from the 16th to the 20th centuries up to 1974 (the year of the Portuguese revolution). The corpus will be a sample of 4,000 letters (2,000 per project) from a universe of millions of epistolary originals.

CARDS's letters were kept unpublished and unknown as material proof inside judicial records, written by people from all social backgrounds (Gomes, Tavares & Guilherme, 2010). FLY's were written in the contexts of war, migration, imprisonment and exile, at a time when members of all social classes were gradually becoming literate, and acquiring writing skills, a process which allowed, eventually, for writing to be adopted by the masses (Alves, 2004) .

1.1 The importance of letters

Documents of the letter genre, and private letters above all, are the best possible data for studying ordinary people in society, their linguistic knowledge and behaviour, as well as their social inscription (Nevalainen, 2004; Hakanen & Koskinen, 2009). Although they are written matter, letters are very close to the informal tenor of spoken utterances. They are frail "light" papers, containing temporary messages, so they seldom reach the dignity of print (publishers only invest in letters by well-known writers, intellectuals or politicians). Kept either in private hands or in archives that filed them for accidental reasons, they are absent from the type of massive linguistic corpora where researchers traditionally go to test hypotheses.

Being private documents written by ordinary people from all social environments – the letters that make up this project make it a relevant tool for various interdisciplinary areas within the Humanities and the Social Sciences, such as Linguistics, Sociology and

History (all the documents edited will be accompanied by indexation and contextualization concerning these three disciplines).

2. How we do it

2.1 XML database and TEI methodology

Having in mind both a standardized method for the editing of manuscripts and the online availability, our team adopted an electronic method using XML language and TEI mark-up conventions designed for editing text. TEI stands for Text Encoding Initiative, a well-known consortium that has worked together since the 1980s and makes guidelines available[1] for primary sources text encoding in the Humanities. This methodology not only honours the manuscript's originality but also allows the proper study and searchability of its contents. The model conceived by TEI uses XML language. XML (Extensible Markup Language) is widely used in the World Wide Web because of the processing and encoding advantages it offers. TEI-XML edition method generates hierarchically organized text files which are machine-readable and people-readable, and thus can serve several purposes such as the editing of a dictionary, transcription of interviews for Oral History or philological annotation of epistolary documents. Above all, the four most important advantages of this method for private manuscripts are: 1) it conserves most features of the manuscript and its calligraphic character, 2) it can be developed and updated while it is being worked; 3) it allows automatic comparison between texts; and 4) the collection transcribed is always searchable as it is written in simple text (and this way no information or any part of any text is lost, resulting in a wide range of opportunities for investigation for every area of interest).

Each XML file has to obey to a DTD, a Document Type Definition[2] , which is a file that states the rules for element-tags and attribute-tags and to a stylesheet, which transforms the tagged file into an edited text.

[1] http://www.tei-c.org/Guidelines/P5/

[2] Our DTD was built after project Digital Archive of the Letters in Flanders (DALF): http://www.kantl.be/ctb/project/dalf/index.htm

2.2 CARDS-FLY methodology

Our database is made up of three files:

(i) XML file for each letter (each letter has a unique code, e.g. CARDS0001 or FLY0001);

(ii) a multimodal presentation: text and image (digital image of the letter's facsimile in JPEG format – each image bearing a unique name, which corresponds to the name of the letter it belongs to);

(iii) a biographic register of authors and addressees with social information in an XML demographic database (each participant has a unique ID so that all the letters they write or receive show their link).

Each letter has two main parts:

a) The header – this contains all the metadata about the document: people involved in the project, funder, research centre, archival information, participants' information (author and addressee), keywords, extra-linguistic context, a concise summary of the letter's contents, project's description and transcription guidelines, among others.

b) The text: it contains the semi-palaeographic edition of the letter. As for letter parts, we consider these:

1. the opener, containing the opening elements such as address, date and salutation;
2. the letter's body (comprising formulaic parts and other tagged text parts, such as abbreviations, deletions or idioms);
3. closer, containing the closing elements such as address, date, salutation, signature and *Post Scriptum*.

Moreover, each letter passes through three steps. Let's look at the example of a postcard sent from an imprisoned Portuguese soldier in a German WW1 prisoner-of-war camp:

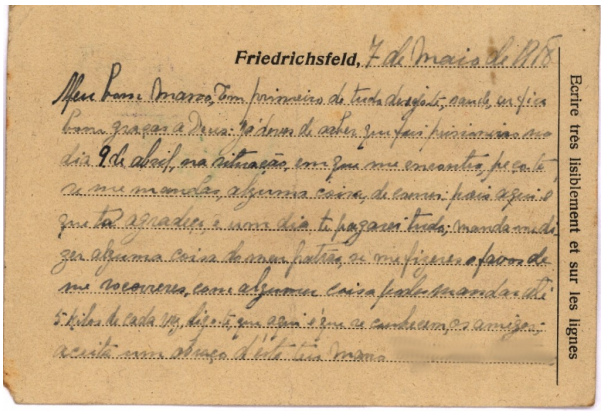

Figure 1: the manuscript

Figure 2: the semi-palaeographic edition in XML

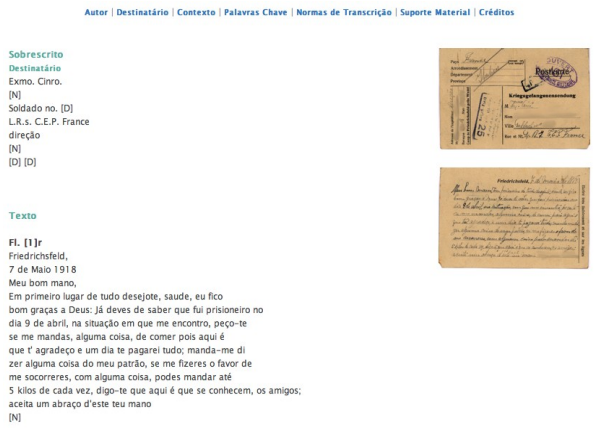

Figure 3: online presentation

Before the FLY letters are published online, we subject them to a special treatment in order to keep private data out of the public scrutiny. All person names, place names (names for small places) and other concrete references are erased from the facsimiles by means of edit image applications. On the other hand, we also suppress those references from the transcriptions, replacing person names by N, place names by L, and other concrete data by D.

3. Results

3.1 Our website

The address: http://www.alfclul.clul.ul.pt/cards-fly

We defined as imperative a presentation of a global introduction to the website, a list of letters (showing its participants, date and a summary of each), a tutorial, as well as downloadable documents that were used for the constitution of the corpus (i.e. stylesheet, DTD, demographic database, DTD of the demographic database, user guidelines and all the letters in XML or .pdf formats). Then, one of the most significant features the website has is a search tool, which guarantees that a wide range of linguistic and historical data can be traced. Users can search by: keywords (provided by specialists in Linguistics, Culture, History and Sociology (the latter for 20th century letters only), year, the name of the author or addressee, and free search. The search tool puts in practice the ultimate goal of the CARDS and FLY projects: that a user may be creative in his/her search for contextualized uses of language through time.

3.2 Already achieved (March 2012)

As far as results are concerned, we have already transcribed nearly 2,000 letters from the 16th to the 19th centuries and 900 letters that cover all four contexts in the 20th century. This represents a total average of 920, 000 words in the CARDS-FLY corpus.

Century	Number of Letters
16th	12
17th	225
18th	539
19th	1144
20th	900
Total	2820

With these two projects we have gathered original data much more varied than the one usually found in epistolary corpora: our corpora have a considerable number of different writers (and addressees) compared to the total sum of letters. Until today the CARDS-FLY projects have gathered 2820 letters written by 1960

different hands (90% men, 10% women) along five centuries and representing all strata of the social spectrum.

3.3 What to do with this data?

3.3.1 How we see the letters

We see these letters as discourse that is very close to face-to-face interpersonal interaction, and that is why:

a) we connect the "rings" within letter chains (letter to letter);

b) we connect actors involved in the written interaction (between letter participants);

c) we make available, as already mentioned, a free search, so that a researcher can link connecting elements between groups of letters, following personal criteria and research purposes.

Our website is not only a means of making data available but also a means of permanent contact between researchers with the same study interests and also between them and the public at large. Moreover, it also allows a dynamic contact between the project's team and website users, since they can easily contact us for various matters, such as doubts related to the letter's transcription or any suggestion concerning difficult readings.

Adding to this, in this case, reading a letter has associated to its critical edition two kinds of problems that we try to solve: a) the difficulty of reading non standard language from different centuries, countries or different codes of writing; this is why we will offer all letters not only its conservative transcription, but also a modernized edition, and a translation into English; b) the difficulty of finding this kind of documents, whether in public archives or in private collections; this is why we always try to offer a facsimile of the manuscripts, along with their transcription. This way, the original data we use are also there for other researchers to analyse, even if they cannot travel into the archives; on the other hand, we help thus to protect the fragile condition of the letters paper.

3.3.2 What to do with these data?

Three studies have already been performed using our letters as a corpus:

1. An information extraction made by computer engineers isolated the polite formulaic parts of 502 letters and studied the implications of their semantics (Hendrickx, Généreux & Marquilhas, 2011).

2. A keywords comparison of the letters' text with those of two larger Portuguese reference corpora using lexical statistics software (WordSmith Tools and AntConc). One reference corpus contains oral utterances, recorded in a Dialectology archive (Cordial-SIN[3]); the other contains texts of several genres written in Contemporary Portuguese (CRPC[4]). One of the conclusions was that the letters' grammatical lexicon is much closer to the one occurring in oral interaction utterances than to the one in written texts (Marquilhas, in press).

3. An attempt to make an automatic normalization of our paleographic transcriptions using a statistical tool of spelling normalization previously designed for the English language (VARD2) (Hendrickx & Marquilhas, 2012).

4. The future and expected results

In December 2011, the team coordinator Rita Marquilhas won an Advanced Grant from ERC to continue work on Digital Humanities for the study of letters. The project to continue CARDS and FLY will be named «P.S. - Post Scriptum: a digital archive of ordinary writings (Early Modern Portugal and Spain)». This project will pursue the goal of making available a corpus of private letters (7,000 Portuguese and Spanish letters) using the same kind of editing method, but adding morphological and discursive tagging so as to make it a very useful corpus for Historical Socio-pragmatics research (Bergs, 2004).

5. Conclusion

With the CARDS and FLY projects, the scientific community as well as the public at large has access to a totally systematized and philologically reliable electronic online edition of a large sample of personal documents of the Portuguese cultural heritage. The corpus works as a basis for the statistical and qualitative study of interactive discourse by ordinary people in the Early Modern ages in a Southern European environment. As

[3] http://www.clul.ul.pt/en/resources/225-description-cordial-sin-syntax-oriented-corpus-of-portuguese-dialects

[4] http://www.clul.ul.pt/en/component/content/article/91/183-reference-corpus-of-contemporary-portuguese-crpc

for the public at large, it benefits from this enterprise in the sense that it has access to easily readable documents which make clear that cultural heritages embrace many characters and life stories, not only those of the celebrate heroes of national narratives.

6. Acknowledgements

This work is funded by the Portuguese Science Foundation, FCT (Fundação para a Ciência e a Tecnologia): PTDC/CLE-LIN/098393/2008, and by Fundação Calouste Gulbenkian.

7. References

Gomes, M. Tavares, L., Guilherme, A. (2010). "estas minhas limitadas cifras tenham a felicidade de acharem a VMce. desfrutando aquela saúde espiritual e corporal tão feliz como lhe deseja o meu afecto" - Different perspectives on correspondance conventionalities. *Proceedings of the Second International Conference on Corpus Linguistics (CILC 2010)*. A Coruña University, pp. 347-357.

Alves, Maria do Céu Garcia dos Reis Loureiro (2003). Um tempo sob outros tempos : o processo de escolarização no Concelho de Mafra: anos de 1772 a 1896. Master's Thesis on Education, Instituto de Educação e Psicologia, Universidade do Minho, (complete work available on an online edition: http://repositorium.sdum.uminho.pt/handle/1822/705).

Bergs, Alexander T. (2004). Letters, A new approach to text typology. *Journal of Historical Pragmatics,* 5:2, pp. 207–227.

Hakanen, M. , Koskinen, U. (2009). From "friends" to "patrons", Transformations in the social power structure as reflected in the rhetoric of personal letters in sixteenth - and seventeenth-century Sweden. *Journal of Historical Pragmatics*, 10:1, pp. 1–22.

Hendrickx, I. Marquilhas, R. (2012). From old texts to modern spellings: an experiment in automatic normalisation. *Proceedings of the Workshop on Annotation of Corpora for Research in the Humanities*, Heidelberg University, Germany, pp. 1-12.

Hendrickx, I., Généreux, M., Marquilhas, R. (2011). Automatic Pragmatic Text Segmentation of Historical Letters. In Sporleder, C., Van den Bosch, A., Zervanou, K. (Eds.), *Language Technology for Cultural Heritage*. Selected Papers from the LaTech Workshop Series. Berlin & Heidelberg: Springer-Verlag, pp. 135-153.

Nevalainen, Tertu (2004). Letter writing. *Journal of Historical Pragmatics,* 5:2, pp. 181–191.