Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Band 5

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr VERLAG

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr
VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

Contents

ANA PAULA BANZA / IRENE RODRIGUES / JOSÉ SAIAS / FILOMENA GONÇALVES

A historical linguistics corpus of Portuguese (16th-19th centuries)

Abstract

The aim of the present paper is to present and discuss a work in progress that involves:

– the creation of online editions of historical documents of a metalinguistic nature, which function both as publications and corpora, allowing for the comparison of manuscript images with the diplomatic edition and providing tools for analysis;

– the application and development of tools that can easily be manipulated by users and adapted to different kinds of historical texts.

The project is still in its first phase, which involves inventorying the metalinguistic texts held by the Évora Public Library (BPE). A survey of the texts of this nature identified in the various catalogues of the library has been carried out. Until now, 43 manuscripts and 200 printed texts with metalinguistic interest, all coming from the reserved catalogues of the BPE, have been identified. In the old reading room catalogue, further 313 works were also identified, while the modern catalogue is yet to be studied. As soon as the inventory is concluded, this will be followed by the organization and the online publication of a catalogue identifying and describing (bibliographical description) the works of a metalinguistic nature held by BPE. The texts' digital processing shall begin after these previous tasks have been completed.

1. Introduction

The Historical Linguistics Corpus (HLC) dealt with in the present paper is a markedly interdisciplinary work, promoting links between linguistics, history and literature on the one hand, and IT on the other hand. It seeks to make available, in an accessible, usable format, a significant number of historical documents of metalinguistic interest, creating tools for the user, and enabling the success of future developments, namely the extension of HLC to cover other works of a similar nature and also works of a different nature.

Providing an online corpus of (meta)linguistic texts of Portuguese between the sixteenth and the nineteenth centuries held by the Évora Public Library, HLC seeks to promote the acquisition of knowledge of some of the most important metalinguistic sources of Portuguese and foster their study, thereby contribut-

ing towards creating online resources in the field and thus advances in research into the language and its history at a national and international level.

These goals are of great importance in view of the current state of knowledge. Although there is a long written tradition of linguistic research in the field of Portuguese, one of the most widely spoken languages in the world, which enjoyed global status even before the advent of the phenomenon of globalization, not enough data or resources have been produced or made available. As a result of this work, we foresee a breakthrough in terms of the acquisition of knowledge on the metalinguistic memory of Portuguese, merging the philological tradition with technical innovation in methodological terms and making a wide range of material, which has been ignored because it is unpublished or rare, available for the first time to researchers and the general public.

The importance of and interest in the dissemination of these texts is bound up with the fact that Portuguese linguistic heritage lies in an indeterminate number of manuscripts and printed texts that often languish undiscovered in libraries and archives. There is a need for the inventorying, systematic processing and publication of such documents. It is known that much of the memory of the Portuguese language has yet to be established precisely due to the absence of corpora that make different text types representing different eras available to researchers all over the world in easily accessible formats. Despite the difficulties inherent in the construction of textual corpora, there is a great need to begin this task, since it is of crucial importance for the survey and analysis of the sources that advances in the study of the linguistic memory be made.

The choice of documents of metalinguistic nature is supported firstly by the fact that among the few corpora which exist for Portuguese there are very few that include texts of this nature. Moreover, the metalinguistic texts for Portuguese from the sixteenth century are recognized as sources with a dual interest for the history of the Portuguese language since, in addition to describing a certain historical state of the language, they are also an example of this state, in this way acting simultaneously as primary and secondary sources.

The proposal to publish documents held by the Évora Public Library is linked with the geographical proximity of the researchers, teachers at the University of Évora, and their role in providing a service to the local community. In addition, the Évora Public Library collection, in spite of its immense value, is difficult to access for researchers, so the benefits of this project go far beyond

the local community, and take on a level of importance at the national and international level.

2. The corpus

We aim to select, from the vast range of material available in the valuable collection held by the library, the works that are considered to be of the greatest value in terms of the "linguistic memory", many of which are unique or rare, and almost all of them poorly catalogued. The project work will first involve the preliminary cataloguing of documents and works with metalinguistic interest, then texts will be selected for publication from among those that are unpublished or whose publications are incomplete and/or not easily accessible.

The criteria used for selection of the texts will be, in addition to the interest they arouse as (meta)linguistic documents, their rarity and/or the fact that they are not available in other corpora, as well as their state of preservation. Below we list only a few of the titles, manuscripts and printed editions that have already been identified and from which will be selected the texts that constitute the HLC, without prejudice to its enlargement in future developmental work:

Manuscripts:

- Apontamentos de orthographia.
- Arte da gramatica e orthographia portugueza, distinta da latina e qualquer outra língua. Dedicada ao real collegio das artes (1600?).
- Castro, P. João Baptista de: Aparato para a Rhetorica, ou Homem Rhetorico.
- Freire, Francisco José (1768): Reflexões Sobre a Lingua Portugueza, Escriptas por Francisco Joze Freire da Congregaçam do Oratorio de Lisboa em 1768. (It should be noted that although the 1842 edition is already available in the online BN version, Freire's original manuscript is in the Évora Public Library.)
- Lima, José dos Santos Baptista e (1740?): Conclusões grammaticaes, dedicadas ao Príncipe D. José por ... Professor em Macau.
- Novo methodo de grammatica portugueza, composto e offerecido ao Exmª Sr. D. Thomás de Almeida, director Geral dos estudos, etc., por João Pi-

nehiro Freire da Cunha, professor de grammatica Latina, n'esta Corte, e natural da mesma.

– Observações do Dr. Pedro José Esteves á Orthographia Portugueza.
– Regras da orthographia portugueza.
– Vocabulario da Letra A.

Printed editions:

– Caldas Aulete, Francisco Júlio (1870): Grammatica Nacional Elementar, adoptada pelo Conselho Geral de Instrucção Publica, Additada com os elementos da língua Concani por J. M. Dias, conforme 3ª ed., Orlim, Na Typographia da India Portugueza.
– Cunha, João Pinheiro Freire da (1770): Breve tratado da orthographia para os que não os estudos ou diálogos ... Lisboa.
– Espada, João Chrysostomo Vallejo (1861): Grammatica portugueza, Lisboa.
– Fonseca, Roque da (1869): Compendio da Orthographia da Lingua Portuguesa, 2ª ed. Correcta e Augmentada com a Orthographia de princípios e varias notas, Margão, Na Typographia do Ultramar.
– Gouveia, J. F. De (1867): Noções Geraes e Elementares de Grammatica Portugueza, Adaptada na Escola Portugueza de Barettos em Cavel, Bombaim, Impressa na Typ. de Viegas & Son.
– Leal, Bento de Araújo (1734): Miscellanea gramatical. Na qual se explicam as partes da oração com todas as suas etymologias, e circunstancias (...), Lisboa, Off. Pedro Ferreira.
– Macedo, José de (com o pseud. de António de Melo da Fonseca): Antidoto da Lingua Portugueza, offerecido ao mui poderoso rei D. João V, Nosso senhor, Amsterdam, em Casa de Miguel Dias (sem indicação do ano, porém a dedicatória é de 1710).
– Pereira, Bento (1655): Florilegio dos modos de fallar e adagios da lingoa portuguesa (...), Lisboa, Por Paulo Craesbeeck.
– Pereira, Pe. José Filipe (1865): Compendio da Grammatica Elementar da Lingua Portuguesa por Systema Philosophico, para uso dos Alumnos das Escholas de Ensino Primario, Orlim, Na Typ. da Ind. Portugueza.

These texts will initially be processed in a conventional manner and will be read and transcribed (in terms of a diplomatic rendering), accurately repro-

ducing the evidence available and preserving all their relevant features (errors, omissions, spelling, word boundaries, abbreviations, etc.). In subsequent stages of the project, the texts will be digitally processed and made available online.

3. Document processing

The digital processing of documents will include the creation and use of resources and tools for natural language processing in order to obtain:

– a document text in an ASCII-like format to enable content analysis;
– electronic dictionaries that can be associated with the documents due to their specific vocabulary;
– the tagging of the ASCII text documents with part-of-speech (POS) markers. These markers enable linguistic researchers to look for word categories in their document analysis;
– the tagging of the ASCII text documents with named entities. These markers can help researchers to look for named entities across the text;
– the tagging of the ASCII documents with a view to sentence polarity. Using sentiment analysis techniques, the sentences of the documents are marked in order to enable researchers to search for sentences where the author's opinion is positive or negative with respect to a particular subject.

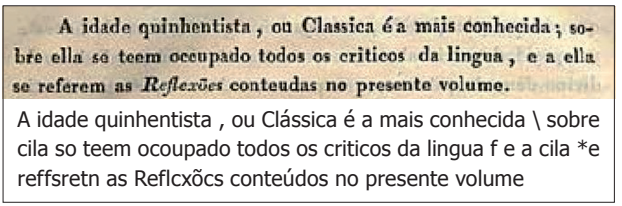
4. Content analysis

The process begins with the application of a text recognition system. Each book is carefully scanned. Then, for each page, the system performs a segmentation of the text areas to be analyzed, typically corresponding to paragraphs. The images corresponding to these areas are converted into text, which can subsequently be treated by natural language processing tools.

The conversion of each image to the text it contains is based on Optical Character Recognition (OCR). We have chosen the analytic OCR approach, trying to identify the individual graphemes, and then make the best interpretation of their sequence. For this interpretation, the system uses a dictionary of terms that are contemporary with the time of each book. As in previous works (Boschetti et al. 2009), our system searches for the best results by combining

the output of more than one OCR tool, such as OCRopus,[1] tesseract,[2] or Abbyy FineReader.[3]

The process is semi-automatic, as successfully performed by other projects-such as PaRADIIT[4] (Ramel/Sidere 2011). The automatic recognition is complemented with human intervention to correct persistent errors. Let's consider an example is the extract below from a book (Freire 1842), with the original being shown at the top of Figure 1 and the recognition result below.


A idade quinhentista , ou Classica é a mais conhecida ; sobre ella se teem occupado todos os criticos da lingua , e a ella se referem as *Reflexões* conteudas no presente volume.

A idade quinhentista , ou Clássica é a mais conhecida \ sobre cila so teem ocoupado todos os criticos da lingua f e a cila *e reffsretn as Reflcxõcs conteúdos no presente volume

Figure 1: Text image and its automatic recognition

One of the mistakes concerns the word *ella* (feminine pronoun, third person, singular, corresponding to English *she*), appearing as *cila* in the OCRed text. Due to the poor image quality, some characters are difficult to recognize. Furthermore, the system dictionary does not contain all the words of the time in question, as is this case with *ella*. In our approach, a supporting dictionary is conceived for each time span, open to receive new terms that are encountered throughout the process of text recognition. In this semi-automatic process, a linguist marks the term *ella* as a valid word for the language of the time, and subsequent occurrences will be correctly identified.

5. Resources and tools

5.1 Electronic dictionary

The electronic dictionary will contain all the terms occurring in the HLC, as well as an ontology featuring the names of the entities mentioned in the texts and their classification.

---

1 http://code.google.com/p/ocropus/. All URLs have been checked and found valid as of late January 2015.

2 http://code.google.com/p/tesseract-ocr/.

3 http://finereader.abbyy.com/.

4 https://sites.google.com/site/paradiitproject/.

The dictionary will primarily be used to support the semi-automatic task of OCR as outlined above. A linguist assigned to this task will accept the dictionary suggestions or add new terms to the dictionary. The interface of the dictionary is integrated with the OCR tools.

The dictionary should contain the morpho-syntactic categories for each word in the corpus as well as other important information such as the current orthography of the word and the list of corpus documents where it occurs.

To give an example: *ella* is 'she' in Old Portuguese orthography; today it is written *ela*. The dictionary entry will look as follows:

> *Ella* --- pronoun, feminine, 3rd person, singular, current: *ela*, used: <list of documents>
>
> *Elle* --- pronoun, masculine, 3rd person, singular, current: *ele*, used: <list of documents>

Later the electronic dictionary will be made available to linguistic researchers both as an object of study and as a tool to support queries in the corpus documents.

The dictionary structure and implementation is an important issue (Hockey 2000, esp. p. 146-171). We use the WordNet structure in our electronic dictionary (Tasovac 2009).

5.2 Part-of-speech tagging

The documents tagged with part-of-speech markers can be used by linguistic researchers for (meta)linguistic analyses such as counting the use of the definite article before a possessive pronoun. This sort of analysis is important to infer the evolution of language phenomena (e.g.: *seus nomes* > *os seus nomes*).

The part-of-speech markers include a large set derived from the Brown corpus currently used for English, enlarged with Portuguese-specific categories which for some analyses can be grouped together into more basic categories such as prepositions, nouns, verbs, etc.

The part-of-speech tagging is a semi-automatic process supervised by a linguist that has to perform tasks such as deciding on the word category markers suggested by the POS-tagger and correcting the tagging of the documents. Since we do not have a training corpus, we use an unsupervised POS-tagger

that is able to infer groups of word categories (Collobert et al. 2011). These groups of word categories correspond to their role in the sentences; e.g., it will group all pronouns in a group and the linguistic supervisor must determine the groups and decide whether to include or exclude certain words in the groups.

5.3 Named entity recognition

This natural language task will give rise to a set of markers in the HLC documents where each name is tagged according to the category it pertains to, such as geographic place, person, or institution. These markers can help researchers, e.g., to count an author's citations throughout a given document in order to infer the impact of that author as a well accepted authority.

The named entity recognition process uses some mixed techniques including machine learning as well as part-of-speech markers and syntactic information.

5.4 Sentiment analysis

Sentiment analysis is a natural language task that uses the results of the part-of-speech tagger and some machine learning techniques in order to infer the topics of the sentences in the document and their polarity. This feature enables the researchers to search for an author's opinions on certain topics. The (meta)linguistic analysis of the author's recommendation for using or not using a given construction can be facilitated with this feature.

6. Conclusions

Besides making a considerable number of meta-linguistic texts dating from the 16th to the 19th century available online, the project aims at the development of tools (which can be made available to others later), beyond the usage of existing tools that allow for the manipulation of data from written texts from several historical periods. This service is really useful for linguists because it facilitates the constitution of specialized corpora and the location of certain words or syntactical structures within them, which is useful for statistical purposes, for example in the study of the diachronic evolution of a given linguistic phenomenon.

References

Boschetti, Federico/Romanello, Matteo/Babeu, Alison/Bamman, David/Crane, Gregory (2009): Improving OCR accuracy for classical critical editions. In: Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL'09). Berlin/Heidelberg: Springer, 156-167.

Collobert, Ronan/Weston, Jason/Bottou, Léon/Karlen, Michael/Kavukcuoglu, Koray/Kuksa, Pavel (2011): Natural language processing (almost) from scratch. In: Journal of Machine Learning Research 12, 2461-2505.

Freire, Francisco José (1842): Reflexões sobre a Lingua Portugueza, escriptas por Francisco José Freire, publicada com algumas annotações pela Sociedade Propagadora dos Conhecimentos Uteis. Lisboa: Typographia da Sociedade Propagadora de Conhecimentos Uteis. http://purl.pt/135.

Hockey, Susan M. (2000): Electronic texts in the humanities: principles and practice. Oxford/New York: Oxford University Press.

Ramel, Jean-Yves/Sidere, Nicolas (2011): Interactive indexation and transcription of historical printed books. Presentation given at: Digital Humanities 2011 (DH2011): June 19-22, Stanford University (USA).

Tasovac, Toma (2009): More or less than a dictionary: WordNet as a model for Serbian L2 lexicography. In: Infoteka – Journal of Informatics and Librarianship 10: 13a-22a.