Building a historical corpus for Classical Portuguese: some technological aspects

Maria Clara Paixão de Sousa*, Thorsten Trippel†

*Instituto de Estudos da Linguagem,
Universidade Estadual de Campinas,
mclara@iel.unicamp.br

†Fakultät für Linguistik und Literaturwissenschaft,
Universität Bielefeld,
thorsten.trippel@uni-bielefeld.de

Abstract

This paper describes the restructuring process of a large corpus of historical documents and the system architecture that is used for accessing it. The initial challenge of this process was to get the most out of existing material, normalizing the legacy markup and harvesting the inherent information using widely available standards. This resulted in a conceptual and technical restructuring of the formerly existing corpus. The development of the standardized markup and techniques allowed the inclusion of important new materials, such as original 16th and 17th century prints and manuscripts; and enlarged the potential user groups. On the technological side, we were grounded on the premise that open standards are the best way of making sure that the resources will be accessible even after years in an archive. This is a welcomed result in view of the additional consequence of the remodeled corpus concept: it serves as a repository for important historical documents, some of which had been preserved for 500 years in paper format. This very rich material can from now on be handled freely for linguistic research goals.

1. Overview

The Tycho Brahe Annotated Corpus of Historical Portuguese (*cf. Corpus Histórico do Português Tycho Brahe* in the references) contains texts written by Portuguese authors born between the 16th and 19th centuries, and represents the largest digital collection of Classical Portuguese documents available today. In the process of building this corpus, which started in 1998, a number of challenges had to be faced – mainly, related to the multiplicity of potential uses for the material.

This paper describes the conceptual and technical restructuring that has been implemented in the corpus since 2005. Our primary goal in this process has been to normalize the documents markup and facilitate information harvesting. Ultimately, the restructuring has enlarged the potential uses of the corpus by different user groups. The requirements of the different text types and user groups were explored and accommodated using transformation functions.

The corpus is available in a valid and hence well formed XML format. The whole system is built exclusively on freely available tools and (proposed) standards such as XSLT (*cf.* Clark 1999) and XQuery (*cf.* Boag et al., 2005), implemented with web interfaces.

2. Motivation

2.1. Background

Between 1998 and 2004, 42 texts (with a total of 1.851.619 words) were included in the Historical Corpus. This collection was intended primarily as material for linguistic studies, during Phase I of the project Rhythmic Patterns, Parameter Setting and Language Change (*cf. Rhythmic Patterns Parameter Setting and Language Change*). The texts were digitalized and annotated according to an initial framework of procedures (*cf.* Britto and Finger, 1999) which targeted strictly at the aims of this linguistic research. The central goal in processing the texts during this phase was to prepare them for analysis by automated tools (Part of Speech Tagger, and Syntactic Parser, *cf.* Finger, 1998). After being scanned and reviewed, the texts were marked up in order to adapt information that could not be handled by the automated tools (such as missing punctuation), or remove information that was not relevant for the tools' analysis (such as headings, page numbers, former editor's comments, etc.). The prepared texts were then processed by the tagging and parser tools. The HTML files, with very simple and non-standardized headings, were stored and listed on a server; no further search or information mining processes were developed then.

The original system was adequate to the initial goal of producing a great volume of linguistic data with precision and rapidity. However, as the research progressed, there came the need for a complete restructuring of the markup and information mining system, for the reasons we expose below.

2.2. Initial Challenges

The restructuring of the corpus was motivated by two central factors: the plan to increase the volume and expand the diversity of texts in the corpus; and the widening of the corpus' potential purposes and public.

The diversity of texts posed a challenge to the existing text structure markup techniques. Some original 16th-17th century editions selected for inclusion presented widely variable spellings, imposing a hindrance to the processing of texts by the part-of-speech tagger, which is partially lexicon-based. However, the immense value such of texts for historical studies could not be neglected.

The increase in the volume of texts posed a challenge for the information mining system. Added to the necessity of continuous update in the corpus, this made it infeasible to continue using a static catalog of texts for accessing the data.

The widening of potential purposes and public posed conceptual challenges for the corpus as a whole. Along the years, the public availability of the historical material in electronic format had raised the interest of different user groups (such as literature scholars, historians, philologists). The access needs of such groups were not well attended to, either by the corpus general presentation, or by the search systems available - both of them being designed for machine processing, not human reading or accessibility.

Our initial challenge was clear then: to accommodate the volume and diversity of texts and the public interest in the corpus. These were our goals:

- to develop a markup system that would allow the inclusion of original prints and manuscripts, which could be used by the automated tools;
- to develop a global information encoding system that would allow agile information mining;
- to develop a user's interface that would be adequate for humans and machines.

In order to achieve these goals, we needed to choose an annotation standard, as we explain below.

2.3. Motivation for the use of standards

Open standards are assumed to be the best way of making sure that resources will be accessible even after years in an archive. Standards and their use tends to become a new buzzword, but the background of maintaining and archiving resources is getting more and more into the focus. Bird and Simons (2003), for example, point out that there are a number of requirements necessary for long lasting accessibility; in working with large selections of historic texts, this becomes even more evident. The content of the texts should be preserved, and the standard markup also needs preservation and documentation.

Considering, then, that open standards are relevant both for the data itself and for the tools involved, our goal was to stick to standardized methods and technology.

The standard in the sense of best practice and agreement is the use of XML; hence the restriction of this corpus project on XML technology, using (lossless) legacy data integration, providing tools for accessing the corpus using the technology developed and standardized for this syntax. A concrete markup schema, the document grammar, is used consistently for the whole corpus; and the intention was to reuse existing proposed markup schemes. The grounds for the selection of the annotation standard will be detailed in the next section.

3. Selecting the annotation standard

The annotation process was based on the following premises:

- the reuse of the available information from the original data collections and the lossless conversion into a more acceptable format;
- possible automatic conversion with a limited amount of additional manual annotation;
- the usability of an annotation schema by linguists;
- possible tools.

A number of annotation schemes were looked at, including TEI, XCES, DocBook:

TEI: the highly recognized suite of schemas is very complex and hardly usable by linguists without a strong background in automatic annotation. Annotation tools were not available besides the usual XML editors such as Oxygen or Emacs. This basically means that the annotators have to work on the code base of the resource. For our annotators (linguists, not interested in code development, but simply in creating marked up resources), schemas more complex than simple HTML (i.e. without the full possible complexity of layout oriented HTML) are not an option. Another problem was the specific information already available in the source, which could not losslesly be mapped onto TEI without adding information manually. An example for this are a number of adaptations of the text for automatic processing, including resegmentation of word boundaries, orthographic modernizations, spelling out of abreviations, etc. Sperberg-McQueen and Burnard (2002, Section 18: Transcription of Primary Sources) provides tags for all of these; but the source data only provided tags for the adapted and the source versions, without indicating the type of change involved. Without massive manual addition of information the TEI could not be used here.

(X)CES: the results of CES have been used to build and develop related standards, such as TEI and the underlying markup schema for TUSNELDA, but the document grammar was not established as a standard, though the name may suggest it. As TEI and (X)CES are related, they share the most common problems of not having an easy-to-use annotation tool available, and of not having a one-to-one relation to the available information in this archive.

DOCBOOK: Though DocBook (*cf.* Walsh, 1999) and later) has a powerfull advocate in its import and export functionalities (for example in *OpenOffice.org*'s office suite), this format is not intended for the markup of corpora, but of documents to be published, archived and automatically processed.

Similar problems occurred in the encoding of metadata which was available in the original idiosyncratic headers (*cf.* Section 2.1). The goal was to use a Dublin Core based header, possibly using the OLAC metadata set for linguistic tools to use the data.

The resulting annotating schema serves our initial purposes of dealing with variable spellings and providing better accessibility. A system for controlled editing was developed, whereby the original spelling can remain available while normalized spelling is added; each format can be made readily available for the relevant purposes. This reaches our goal of maintaining important philological information while at the same time providing clean, normalized material for automated analysis of part-of-speech and syntax. Concerning the accessibility challenges, the original, weakly systematized metadata was restructured using the OLAC metadata set (*cf.* Simons and Bird, 2002) filled with information recorded with the corpus. The creation of the catalog relies on some metadata according to the OLAC standard.

4. Document supply chain

Figure 1 shows the design of the necessary system for supplying the corpus information. The corpus itself can be stored on the a server, which is accessible from the web.

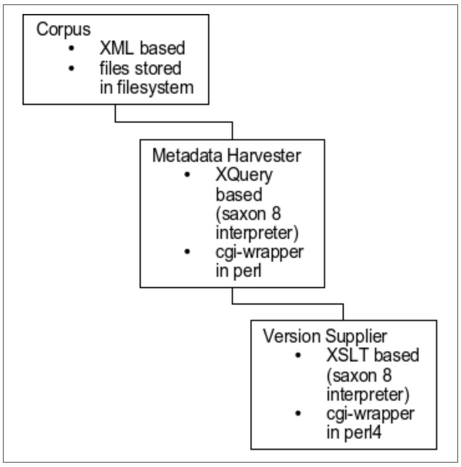

Corpus
• XML based
• files stored in filesystem
Metadata Harvester
• XQuery based (saxon 8 interpreter)
• cgi-wrapper in perl
Version Supplier
• XSLT based (saxon 8 interpreter)
• cgi-wrapper in perl4

Figure 1: System design for supplying a corpus

The corpus files are harvested for the metadata to allow the users to select a subset of the whole corpus, such as a single text or a text by one specific author. For this process, the Metadata Harvester can use XQuery as the standard to query large amounts of XML coded data.

By selecting a desired output version the selected subcorpus can be transformed into this version. The method of choice for the transformation of complete XML documents is XSLT.

5. Preparing the documents

The process of document preparation as indicated in the previous section consists of three major phases, the corpus preparation, the metadata harvesting and the document delivery. Each will be explored in the following.

5.1. Document preparation

The texts are marked up in XML to be prepared for the stages of metadata harvesting and document delivery. The restructuring process included two different lines of action in this aspect: the development of the markup itself (*cf.* 5.1.1 below); and the adaptation of the existing documents to this markup (*cf.* 5.1.2 below).

The resulting system had to be flexible. At the present stage, additional corpus data is frequently added, and some updates may be made after the original launch of the archive. Therefore, the applications working on this, now normalized data, needed to be dynamic on demand, i.e. working on the data while providing the required information. With this method, the users can rely on getting the most up-to-date version of the corpus in process, before a code freeze is to be expected allowing proper versioning.

5.1.1 Adaptation of the former markup

Starting with the existing pseudo-normalized but machine readable corpus version (the 42 original texts), a normalized version had to be prepared. This was done by a semi-automatic process including a Perl-script translating the existing markup into XML markup by translating element names and adding attributes and end tags as appropriate. It also included simple algorithms to disambiguate overlapping changes, which were the results from word-resegmentation. The result needed to be manually checked with the assistance of a syntactic parser. A similar process was used to prepare well-formed metadata. However, as the content was a bit more restricted, the manual checking of the metadata was easier and simpler to complete.

5.1.2 Preparation of new materials

After the development of the markup standards, all the material to be introduced in the corpus now follows the new preparation system from the start. The files are transcribed and receive the initial markup for metadata and text-structure (paragraphing, sentencing, etc.).

In a second stage, the texts go through an edition process. One of the central goals of the whole restructuring process, as mentioned in section 2, was to allow the inclusion of original 1500-1700 prints with variable spellings. The spelling variations are normalized in the transcribed texts, and all interferences (editions) are encoded in XML. *Figure 2* shows the edited markup of the modern word *alguma* in one of the texts according to the markup scheme (*cf.* Paixao de Sousa, 2005):

```
<v id=”g_008_v_382” type=”mod”>
     <ed id=”g_008_ed_382”>alguma</ed>
     <or id=”g_008_or_382”>algũa</or>
</v>
```

Figure 2: Edition markup: Original word: “algũa” ('some'), modern word: “alguma”

This edition system has three main advantages:

- it allows the editors to keep track of their interference in the texts;
- it maintains the integrity of the original writing;
- it permits different users to access adequate versions.

5.2. Metadata harvester

Different filters for the OLAC metadata were pre-defined to help users to find reasonable subsets of the large corpus. The access to the corpus data is automatized without human interference and according to the metadata that has been defined before. The material is available on the web for the community that is spread throughout the world; the corpus is continually extended and updated providing for a more systematized description as it becomes available.

The resulting catalog is based on different criteria, i.e. access by the name of an author, genre, or time period. The method of choice again was using the corpus inherent information to create the catalog, using XQuery.

Many different filters to the corpus are predefined, adjusted to the domain of this special corpus. They include:

- author: picklist of texts by author's name
- title: picklist of texts by title
- list by genre
- chronological list

As these lists are generated, additional filters according to all recorded metadata categories can be defined. The reason for this is that the metadata model is based on attribute-value form, hence the list of available categories can be used to create the possible filters.

5.3. Version supplier

The different user groups required different views on the data, and the different views were to be created automatically from the source to allow updates in the source taking effect in the served versions. Hence the views had to be implemented as a transformation from the original, also to be processed when required. The transformation was implemented using XSLT. Currently, we provide the following versions:

- unedited version, based on the transcript of the original spelling (*cf.* Figure 3);
- modernized version with normalized orthography, formated for web-browsers (*cf.* Figure 4);
- modernized version with normalized orthography, formated for non XML enabled legacy tools (*cf.* Figure 5);
- lexicon of editions, with the modernized and the original items in each variant (*cf.* Figure 6)

Additional views can be created based on new XSL stylesheets.

6. Creating the User Interface

6.1. First concept

The user interface design was heavily influenced by the user community, which is distributed and not very likely to be willing to provide and write their own tools. Even the installation of seldomly used tools poses a serious challenge to these potential users, especially if this user interface does not have the "look and feel" of well-known applications. This also excludes the option of having a command line interface. A graphical user interface (GUI) within a well known application was necessary. Plugins to existing software requires installation and imposes restrictions on possible applications.

Considering the context of distributed users, and with the idea of not imposing a complex software infrastructure, we decided to use a central data repository with web based access. On the user side, the client should be a web-browser, not relying on proprietary extensions, but on web standards such as HTML – and maybe, if scripting is needed on the client side – JavaScript. For the access to the database, the available metadata had to be put into web forms for subcorpus selection, not overloading the user interface with programming like queries, but with sufficient guidance for the user to get the information of interest.

The restriction to standard web technology on the client side also resulted in a high measure of portability: the system is accessible independent of computer platform, provided a web-browser is available. Initial tests were conducted on different browsers and different platforms: the Microsoft Internet Explorer 6 on Windows; Safari on Mac OS; Firefox on Linux and Windows; Konqueror on Linux; and also "exotic" platforms such as Palm OS 5.4 with the Blazer browser.

The client side web-browser access also points to the requirement of server side processing, i.e., the querying and transformation of the corpus has to be conducted on the server side. Based on the same principles of portability, standard conformance and finally the Web processing, the XML related technologies were selected, with some simple wrappers in a scripting language, as already indicated by *Figure 1*. These wrappers were necessary to include parameters when calling an interpreter to process transformations and queries. Both wrappers only contain a simple call for the interpreter with the parameters of files and query or transformation functions.

6.2. Problems with the Technology

During the implementation a few problems had to be faced. The first problem was related to providing the data on the first hand. Problems arised due to digitalization problems and unstable orthography, not only in segmentation but also in the use of characters and character encodings. The non-normalized orthography finds its counterpart in a non standardized characters, sometimes distributed systematically and sometimes in free variation. Hence Unicode had to be used, as ISO 8859 encoding did not provide all required characters.

After the normalization of the source a couple of problems where left:

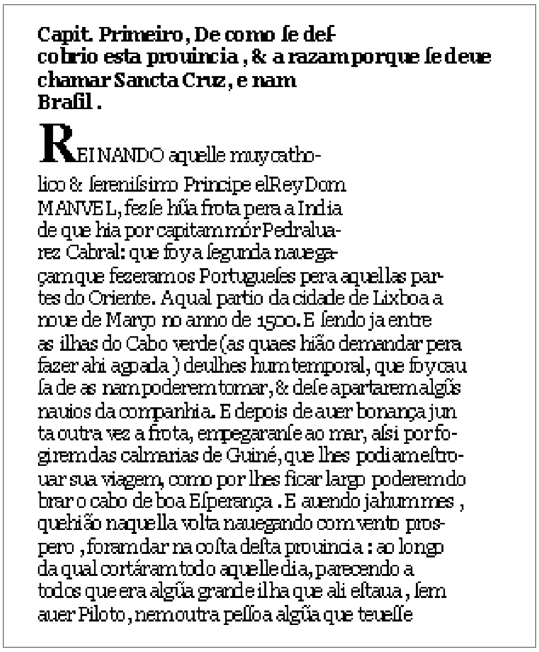
Scalability on the client side: As the database grew and was tested for larger amounts of data using preliminary versions of subcorpora, it became clear that the original GUI design was unsuitable. The original GUI showed lists of subcorpora by author name, era, title of text, etc., which were very user-friendly while looking at 4 preliminary texts. When the number of subcorpora was increased, a list-based user interface became unusable, due to rendering time and amount of information presented on one webpage. The GUI had to undergo redesign, being splitted into different pages according to anticipated uses and hiding the lists by using picklists.

Scalability on the server side: the increasing amount of data also pushed up the processing time recognizably. This could be avoided by either one of two strategies: by optimizing the queries, avoiding redundant read operations on the file systems; or by running the metadata harvester only after adding new additional subcorpora. Both resulted subjectively in acceptable respond times. In case this should prove insufficient when adding even larger amounts of data, the use of an XML database management system will need to be considered; however, this is not intended at the present stage of the project.

6.3. Present Interface

To illustrate the present webinterface of the corpus, we show here screenshots from different versions rendered from one of the documents (*cf.* Tycho Brahe Corpus Document g_008).

Figure 3 shows the version with the original orthography and word segmentation; the layout is included using CSS for HTML:

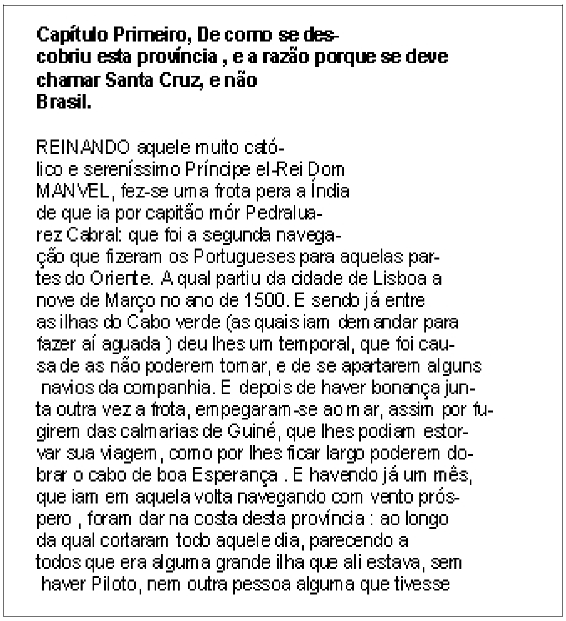


Capit. Primeiro, De como ſe deſ-
cobrio esta prouincia, & a razam porque ſe deue
chamar Sancta Cruz, e nam
Braſil.

REINANDO aquelle muy catho-
lico & ſereniſsimo Principe elRey Dom
MANVEL, fezſe hũa frota pera a India
de que hia por capitam mór Pedralua-
rez Cabral: que foy a ſegunda nauega-
çam que fezeramos Portugueſes pera aquellas par-
tes do Oriente. Aqual partio da cidade de Lixboa a
noue de Março no anno de 1500. E ſendo ja entre
as ilhas do Cabo verde (as quaes hião demandar pera
fazer ahi agoada) deulhes hum temporal, que foy cau
ſa de as nam poderem tomar, & deſe apartarem algũs
nauios da companhia. E depois de auer bonança jun
ta outra vez a frota, empegaranſe ao mar, aſsi por fo-
girem das calmarias de Guiné, que lhes podiam eſtro-
uar sua viagem, como por lhes ficar largo poderem do
brar o cabo de boa Eſperança. E auendo ja hum mes,
que hião naquella volta nauegando com vento pros-
pero, foram dar na coſta deſta prouincia: ao longo
da qual cortáram todo aquelle dia, parecendo a
todos que era algũa grande ilha que ali eſtaua, ſem
auer Piloto, nem outra peſſoa algũa que teueſſe

Figure 3: Version I: original version

Figure 4 shows the same instance, in the normalized orthography:

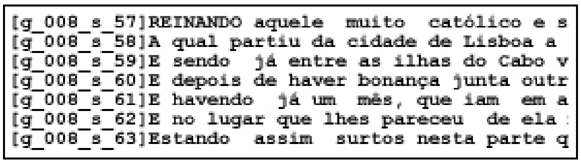


Capítulo Primeiro, De como se des-
cobriu esta província, e a razão porque se deve
chamar Santa Cruz, e não
Brasil.

REINANDO aquele muito cató-
lico e sereníssimo Príncipe el-Rei Dom
MANVEL, fez-se uma frota pera a Índia
de que ia por capitão mór Pedralua-
rez Cabral: que foi a segunda navega-
ção que fizeram os Portugueses para aquelas par-
tes do Oriente. A qual partiu da cidade de Lisboa a
nove de Março no ano de 1500. E sendo já entre
as ilhas do Cabo verde (as quais iam demandar para
fazer aí aguada) deu lhes um temporal, que foi cau-
sa de as não poderem tomar, e de se apartarem alguns
navios da companhia. E depois de haver bonança jun-
ta outra vez a frota, empegaram-se ao mar, assim por fu-
girem das calmarias de Guiné, que lhes podiam estor-
var sua viagem, como por lhes ficar largo poderem do-
brar o cabo de boa Esperança. E havendo já um mês,
que iam em aquela volta navegando com vento prós-
pero, foram dar na costa desta província: ao longo
da qual cortaram todo aquele dia, parecendo a
todos que era alguma grande ilha que ali estava, sem
haver Piloto, nem outra pessoa alguma que tivesse

Figure 4: Version II: normalized orthography

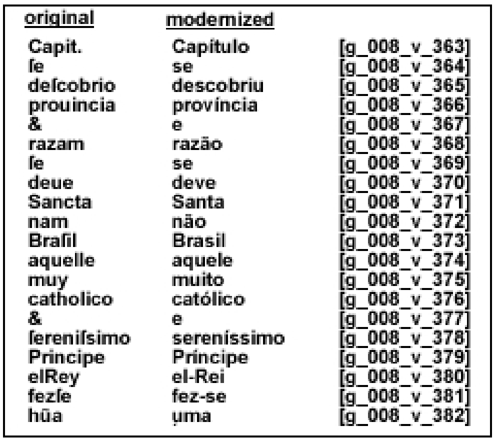
The format for the part-of-speech tagger and the syntactic parser has to be slightly different, text based in a non-XML based format and without any other layout. This version is shown in *Figure 5*:



```
[g_008_s_57]REINANDO aquele  muito  católico e s
[g_008_s_58]A qual partiu da cidade de Lisboa a
[g_008_s_59]E sendo  já entre as ilhas do Cabo v
[g_008_s_60]E depois de haver bonança junta outr
[g_008_s_61]E havendo  já um  mês, que iam  em a
[g_008_s_62]E no lugar que lhes pareceu  de ela
[g_008_s_63]Estando  assim  surtos nesta parte q
```

Figure 5: Version III: normalized version for technical post processing

For historical linguistics studies, it is essential to list the possible variants for each word in the texts. This is easily rendered from this markup; a lexicon of normalized words and their original variant is shown in *Figure 6*. It also shows the identifiers of the relevant segments, which can be used for concordancing the text in the next phase of development.



original	modemized	
Capit.	Capitulo	[g_008_v_363]
ſe	se	[g_008_v_364]
deſcobrio	descobriu	[g_008_v_365]
prouincia	provincia	[g_008_v_366]
&	e	[g_008_v_367]
razam	razão	[g_008_v_368]
ſe	se	[g_008_v_369]
deue	deve	[g_008_v_370]
Sancta	Santa	[g_008_v_371]
nam	não	[g_008_v_372]
Braſil	Brasil	[g_008_v_373]
aquelle	aquele	[g_008_v_374]
muy	muito	[g_008_v_375]
catholico	católico	[g_008_v_376]
&	e	[g_008_v_377]
ſereniſsimo	serenissimo	[g_008_v_378]
Principe	Principe	[g_008_v_379]
elRey	el-Rei	[g_008_v_380]
fezſe	fez-se	[g_008_v_381]
hũa	uma	[g_008_v_382]

Figure 6: Version IV: lexicon of orthographic variation

7. Results so far

The newly restructured corpus, the corpus catalog and the sources are available via <http://www.ime.usp.br/~tycho/corpus>. All the 42 texts from *Phase I* (1998-2004) have been adapted into the XML format, and are included in the catalog; of these, 22 are final releases, and 20 are in the stage of final review of editions markup.

In 2005, four new texts have been prepared for inclusion in the Corpus, following the new markup framework:

i. Gandavo, M.: "*Historia da prouincia de Sancta Cruz a que vulgarmente chamamos Brasil*". (original print: Lisbon, 1576; *cf.* Gandavo, P.M. de)

ii. Galvão, D.: "*Chronica do muito alto e muito esclarecido principe D. Affonso Henriques primeiro Rey de Portugal*" (original print: Lisbon, 1726; *cf.* Galvao, D.)

iii. Lopes, F: "*Chronica del Rey D. Ioam I de Boa Memoria e dos reys de Portugal o decimo*". (original print: Lisbon, 1644; *cf.* Lopes, F.)

iv. Pina, R.: "*Chronica do muito alto e muito esclarecido principe Dom Diniz, sexto rey de Portugal*". (original print: Lisbon, 1729; *cf.* Pina, R. de)

These four important historical chronicles were transcribed from original 16th-17th century prints, via facsimiles recently published by the *Bibilioteca Nacional de Lisboa* in the form of digital pictures (cf. [Biblioteca National Digital]). The original prints' layout and orthography have been faithfully preserved in the transcription. The inclusion of such original prints would have been impossible using the former preparation techniques in the Corpus.

One of these texts (Gandavo, (i) above) has been fully edited, and the resulting version with modernized orthography is now ready to be processed by the automated part-of-speech tagger.

Apart from the prints, we are currently adapting the XML markup to be applied to the transcription of original manuscripts – a material which is even more valuable than the original prints, to historical linguists and philologists.

Philological transcription of ancient manuscript texts traditionally involves complex markup to encode editors' interferences (such as indication of deteriorated material, identification of different handwritings/authors, etc.). This markup, normally done by hand or by loosely structured use of regular text editors, is going to be fully adapted to standardized XML annotation in the next phase of our project. As a preliminary study for the development of a standardized transcription technique for manuscripts, a group of 30 manuscripts written between 1600-1700 has already been transcribed with a test XML annotation, with interesting results. At present, another, larger group of manuscripts (about 200 documents written between 1800-1900) are being adapted.

Beyond the clear advantages for linguistic research, the controlled, standardized transcription of these original prints and manuscripts has a broader consequence: it helps preserve important historical documents, at the same time that it makes them available in a fully public and proprietary-free framework.

8. References

Biblioteca Nacional Digital. URL: <http://bnd.bn.pt>

Bird, S. and G. Simons (2003). Seven dimensions of portability for language documentation and description. Language , 79(3):557–582, September.

Boag, S., D. Chamberlin, M. F. Fernandez, D. Florescu, J. Robie, and J. Siméon (2005). XQuery 1.0: An XML query language. URL: <http://www.w3.org/TR/xquery/>, W3C Candidate Recommendation 3 November 2005

Britto, H. and M. Finger (1999). Constructing a parsed corpus of historical Portuguese. In Proceedings of International Humanities Computing Conference, University of Virginia, Charlottesville. ACH/ALLC.

Clark, J. (1999). XSL Transformations (XSLT) Version 1.0. <http://www.w3.org/TR/xslt>. W3C Recommendation, November 1999.

Corpus Histórico do Português Tycho Brahe. URL: <http://www.ime.usp.br/~ tycho/corpus>.

Finger, M. (1998). Tagging a morphologically rich language. In Proceedings of the first workshop on text, speech and dialogue (TSD), Brno.

Galvao, D. Chronica do muito alto e muito esclarecido principe D. Affonso Henriques primeiro Rey de Portugal/ composta por Duarte Galvão ; fielmente copiada do seu original, que se conserva no Archivo Real da Torre do Tombo... por Miguel Lopes Ferreira. - Lisboa Occidental : na Officina Ferreyriana, 1726. - [23], 95 [1] p. ; 27 cm.

Gandavo, P.M de. História da prouincia Sãcta Cruz que vulgarme[n]te chamamos Brasil/ feita por Pero Magalhães de Gandauo. Em Lisboa: na officina de António Gonsaluez: vendense em casa de Ioão Lopez, 1576. - 48 f. : 1 est. ; 4º (18 cm) - Assin: A-F//8. - Anselmo 709. - Faria - BN Rio de Janeiro p. 38. - B. Museum 150 coln 204.

Ide, N. and Priest-Dorman, G. (2000). Corpus Encoding Standard. <http://www.cs.vassar.edu/CES/>; XML version: <http://www.cs.vassar.edu/XCES/>

Lopes, F. Chronica del Rey D. Ioam I de Boa Memoria e dos reys de Portugal o decimo / composta por Fernam Lopez. Em Lisboa: Antonio Alvarez, 1644. - 2 v.;28 cm. BN H.G. 2551V. BN H.G.2552 V.

Paixao de Sousa, M. C. (2005). Text Annotation Manual for The Tycho Brahe Corpus. URL: <http://www.ime.usp.br/~tycho/corpus/manual/prep/manual_e.html>.

Pina, R. de, (Ferreira, ed. Lit). Chronica do muito alto e muito esclarecido principe Dom Diniz, sexto rey de Portugal / composta por Ruy de Pina.; fielmente copiada do seu original por Miguel Lopes Ferreyra. Lisboa Occidental: Na Off. Ferreyriana, 1729. - [12], 107 p. ;31cm BN H.G. 11683//6 V.

Rhythmic Patterns, Parameter Setting and Language Change. URL: <http://www.ime.usp.br/~tycho/prfpml/english/e_index_2.html>

Sperberg-McQueen, C. M. and Burnard, L. (eds.) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen

Simons, G. and S. Bird (2002). Open Language Archive Community (OLAC) metadata. URL: <http://www.language-archives.org/OLAC/metadata.html>, December.

Tycho Brahe Corpus Document g_008. URL: <http://www.ime.usp.br/~tycho/corpus/texts/xml/g_008.xml>.

Walsh, N. (1999). DocBook: The definitive guide. O'Reilly, Sebastopol