Taming the Tiger Topic: An XCES Compliant Corpus Portal to Generate Subcorpora Based on Automatic Text-Topic Identification

Marcelo Muniz,[1] Fernando V. Paulovich,[1] Rosane Minghim,[1]
Kleber Infante,[1] Fernando Muniz,[1] Renata Vieira[2] and Sandra Aluísio[1]

Abstract

Large-corpus projects generally use a rich header to describe their texts allowing several types of text searching to create study subcorpora. They normally use TEI (Text Encoding Initiative) or XCES (Corpus Encoding Standard for XML) as encoding standards. TEI was a very early initiative on standardizing text encoding. XCES is currently being largely used in corpus-based work in natural language processing (NLP) applications since it allows the use of stand-off annotation. In the PLN-BR project which is being developed under the sponsorship of the funding agency CNPq, Brazil, we also use XCES. Similarly to the ANC (American National Corpus), we use the implementation of the specifications of ISO TC37 SC4's Linguistic Annotation Framework (LAF). In this paper, we present the *Portal de Córpus*[3], an XCES compliant corpus portal, which gives access to several Brazilian Portuguese newspaper corpora compiled in the scope of PLN-BR project. Moreover, we present one of the PLN-BR corpora, named PLN-BR GOLD, which is freely available in the Web. The *Portal de Córpus* is based on a database that maps the XCES header elements into relational entities. The whole framework can be easily used in other projects of Brazilian Portuguese (BP) corpora using the same standard, since we also made available a header editor and corpus uploader to import and edit full XCES-compliant headers. We provide several searching functions to build study corpora from a main corpus. The searches are based on the elements presented on the text headers, such as, bibliographic information, newspaper sections, text types and keywords. While the major part of the information which can be included in the header is based on external text information, topic is one of them which should be recovered based on internal information. This paper also presents our approach to allow easy access to the corpus text topics by providing content-based visual maps of the texts using multidimensional projections. A tool, called Projection Explorer (PEx)[4], developed at ICMC-USP, was adapted to be part of our Portal. *PEx-Corpus Tool* – the adaptation – uses term covariance technique to extract discussed topics within the corpus texts and is activated after a user has created a subcorpus based on several information presented in the text header. With *PEx-Corpus Tool* the user can visually inspect the subcorpus to explore its content and create further subcorpora based on a selection of topics.

[1] University of Sao Paulo, Sao Carlos, SP, Brazil
e-mail: marcelo.muniz@gmail.com, paulovic@icmc usp.br, rminghim@icmc.usp.br, corujito@gmail.com, fernando.muniz@gmail.com, sandra@icmc.usp.br
[2] Universidade do Vale do Rio dos Sinos, Sao Leopoldo, RS, Brazil
e-mail: renata.vieira@gmail.com
[3] http://www.nilc.icmc.usp.br:8180/portal/
[4] http://www.lcad.icmc.usp.br/~paulovic/pex/

1. Introduction

Large-corpus projects, like BNC (British National Corpus) for British English variant, and ANC project for American English, contribute for the description of the English language and for the development of resources, e.g., dictionaries and grammars. What can be less visible at a first analysis is that they also contribute for the development of natural language processing (NLP) tools, such as lemmatizers, POS taggers, parsers, tools for anaphora annotation, which will give support for the linguistic annotation of these corpora. Moreover, they contribute for the development of annotations and encoding formalisms, like TEI (Burnard, 1995) and XCES (Ide et al, 2000; Ide and Romary, 2004), which uses XML as representation format. XCES is currently largely used in corpus-based work in NLP applications, since it allows the use of stand-off annotation. There are several advantages in using stand-off markup[5]: the possibility to have levels of annotation which have crossing branches (not normally possible in XML); new levels of annotation can be added without disturbing existing ones; editing one level of annotation has minimal incidental effects on others; and annotators can work on different levels at the same time without worrying about creating different versions.

The reusability and extensibility are pointed out as two aspects to be taken into account in corpora projects, according to Ide and Brew (2000). As for the text typology, it is expected that texts of large corpora are diversified, for example, in authorship, genre, text types, topics and domain.

According to Biber (1993), the text samples must include all the linguistic variations in the language in order to build a representative corpus for distinct language studies. Biber proposes that corpus work must proceeds in a cyclical fashion, in which a pilot corpus should be compiled first based on external criteria for text selection, followed by empirical investigations of linguistic variation and design revision. This *modus operandi* can determine a modification of several design parameters that will include more texts that will be analyzed again. The point that Biber wants to make with his proposal of cyclical corpus creation is that text type like, for example, letters, articles, news article, manual and report, can not be identified a priori, as they represent the text groups that are similar in their linguistic features. We can say the same for topics that could be better identified if we could take into account linguistic features and not just be based on an external, non-linguistic classification, which in general is not generally accepted. According to Eagles (1996), topic is the lexical aspect of internal analysis of a text.

In the scope of a Brazilian project named PLN-BR, several corpora of newspaper texts were compiled: PLN-BR FULL, PLN-BR CATEG, and PLN_BR GOLD. For the last two corpora, the metadata text type was based on an automatic classification, which used text internal features. This classification used a text type classifier trained on the 40-text types of Lácio-Web Project[6] (Aires et al, 2004; Aires et al, 2005). PLN-BR GOLD corpus, its segmentation and linguistic annotations will be presented in Section 2. With regard to text topic, we present in Section 5 our approach to allow easy access to the corpus text topics, by providing content-based visual maps of the texts using multidimensional projections. A tool, called Projection Explorer (PEx), was adapted to be part of our Portal. *PEx-Corpus Tool* – the adaptation – uses term covariance technique to extract discussed topics within the

[5] http://www.ldc.upenn.edu/annotation/database/presentations/Isard/Standoff_IRCS_Isard.ppt.
[6] http://www.nilc.icmc.usp.br/lacioweb/

corpus texts and is activated after a user has created a subcorpus based on several information presented in the text header. In Section 3, we present *Portal de Córpus*, an XCES compliant corpus portal to give access to Brazilian Portuguese newspaper corpora compiled in the scope of PLN-BR project. The whole framework can be easily used in other BP corpus projects using the same standard, since we also made available a header editor and corpus uploader to import and edit full XCES-compliant headers. The header editor and corpus uploader will be presented in Section 4. Section 6 presents our conclusions and future work.

2. PLN-BR project and the public release of the PLN-BR GOLD corpus

The PLN-BR project – Resources and tools for information retrieval from Portuguese textual bases – funded by CNPq, has as goal the construction of common resources and tools to be shared by a group of Brazilian universities. The project integrates 7 universities which are building and using the same corpus. The larger corpus of newspaper texts related to this project is called PLN-BR FULL. It contains 103.080 thousands texts from the Brazilian newspaper Folha de São Paulo (FSP) with 29.014.089 tokens. From this corpus two other corpora were generated, the PLN-BR CATEG with 30 thousands texts and 9.780.220 tokens which is intended to serve to research on text classification and another one, called PLN-BR GOLD, which has 1024 texts and 338.441 tokens and which is a portion of the corpus that has been distributed together with further linguistic annotation.

The XML encoding used in the PLN-BR project follows those of the ANC project, which is based on XML Corpus Encoding Standard (XCES) both for primary data and further linguistic annotation.

2.1 Primary data encoding

The PLN-BR GOLD corpus is annotated in paragraphs and sentences as well as with external header annotation. Each logical document in this corpus is an XML file that follows the XCES schema xcesDoc.xsd. Physically, the primary data and its annotations are provided in multiple XML documents generating a directed graph referencing regions of primary data. Usually the stand-off format allows flexibility for creators and users of the corpus, however if the user prefers in-line annotation there is also the merged version which unites the logical structure and sentence boundaries annotations in one same XML file. This is done on the basis of the ANC Merge Tool[7]. The logical mark up e sentence boundaries annotations were created using SENTER[8], a sentence segmentation tool for Portuguese texts. Another generated file was the XML header. The following example (Figure 1) is a XML header file of PLN-BR GOLD.

```
<?xml version="1.0" encoding="UTF-8" ?>
<cesHeader xmlns=http://www.xces.org/schema/2003
  xmlns:xlink="http://www.w3.org/1999/xlink"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.xces.org/schema/2003" version="1.0.4">
   <fileDesc>
```

[7] http://www.americannationalcorpus.org/tools/index.html
[8] http://www.icmc.usp.br/~taspardo/Senter.htm

```
    <titleStmt>
      <title>2005ago_751</title>
      <respStmt>
        <respType>Criação do Header</respType>
        <respName type="person">Kleber Infante<respName>
      </respStmt>
      <respStmt>
        <respType>Criação do Header</respType>
        <respName type="person">Marcelo Muniz<respName>
      </respStmt>
    </titleStmt>
    <extent>
      <wordCount>194</wordCount>
      <byteCount units="bytes">2120.0</byteCount>
      <extNote>1</extNote>
    </extent>
    <publicationStmt>
      <pubAddress>Av. Trabalhador São-carlense, 400 - Centro, Caixa Postal:
       668 - CEP: 13560-970 - São Carlos - SP</pubAddress>
      <telephone>+55 16 33739663</telephone>
      <eAddress type="www">http://www.nilc.icmc.usp.br</eAddress>
      <pubDate>2006</pubDate>
    </publicationStmt>
    <sourceDesc>
      <biblStruct>
        <monogr>
          <title>Com três gols, Adriano brilha no Italiano</title>
          <author>DA REPORTAGEM LOCAL</author>
          <respStmt>
            <respType>crédito</respType>
            <respName type="institution">DA REPORTAGEM LOCAL</respName>
          </respStmt>
          <imprint>
            <pubPlace>Folha de São Paulo</pubPlace>
            <publisher type="org">Empresa Folha da Manhã S.A.</publisher>
            <pubDate>Segunda-feira, 29/08/2005</pubDate>
            <pubAddress>São Paulo</pubAddress>
          </imprint>
          <biblNote>ESPORTE</biblNote>
          <biblScope type="PP">D5</biblScope>
        </monogr>
      </biblStruct>
    </sourceDesc>
  </fileDesc>
  <profileDesc>
    <textClass>
      <catRef target="genero.8 genero.8.18 genero.8.18.10 distribuicao.12
       tipotextual.24" />
     <keywords>
       <keyTerm>
         <![CDATA[ DESEMPENHO ]]>
       </keyTerm>
       <keyTerm>
         <![CDATA[ FUTEBOL   ]]>
       </keyTerm>
       <keyTerm>
         <![CDATA[ CLUBE  ]]>
       </keyTerm>
       <keyTerm>
         <![CDATA[ ITÁLIA  ]]>
       </keyTerm>
       <keyTerm>
         <![CDATA[ CAMPEONATO ITALIANO ]]>
       </keyTerm>
       <keyTerm>
         <![CDATA[ INTERNAZIONALE ]]>
       </keyTerm>
       <keyTerm>
```

```
          <![CDATA[ TREVISO ]]>
        </keyTerm>
        <keyTerm>
          <![CDATA[ ADRIANO ]]>
        </keyTerm>
      </keywords>
    </textClass>
    <annotations>
      <annotation type="logical" ann.loc="ESPORTE_2005_760-
        logical.xml">Logical markup</annotation>
      <annotation type="s" ann.loc="ESPORTE_2005_760-s.xml">Sentence
        boundaries</annotation>
      <annotation type="content" ann.loc="ESPORTE_2005_760.txt">Document
        content</annotation>
     </annotations>
  </profileDesc>
</cesHeader>
```

Figure 1: An XML header with information about the text typology used in Lácio-Web project. In this header, the information "genero.8.18.10" is related to the newspaper genre; "distribuicao.12" is related to newspaper as a medium; and "tipotextual.24" is related to news as a text type.

The texts and the logical mark up annotation included 2 blank spaces and a line break in the beginning and in the end of all texts so that the values of text, body, and div will be always 0, 1 e 2, with no overlapping.

2.2 Linguistic annotations

Full syntactic annotation of the corpus is provided on the basis of the Palavras Parser, a multi-level constraint grammar parser described in (Bick, 2002). Previous XML annotations have been proposed for the Palavras parser, as shown in (Gasperin et al, 2003). The current version of the parser produces various output styles, including Tiger XML[9]. On the top of that we generate XCES conformant linguistic annotation, as previously discussed in (Vieira et al 2003) (see Figure 2) to be distributed with the GOLD corpus. Tokens are linked to the main text through the attributes *from* and *to* in the structures of type *token*, for each token there is a corresponding POS structure. Phrases are identified for group of tokens, as seen in the example below for the NP "A universidade", subject noun phrase in the sentence "A universidade precisa de idéias criativas". For the complete tag and feature set provided by Palavras for POS and Phrases, see the VISL project web site[10].

Tokens

```
<struct type="token" from="0" to="1">
  <feat name="id"       value="t1"/>
  <feat name= "base"    value="A"/>
</struct>
<struct type="token" from="2" to="8">
  <feat name="id"       value="t2"/>
  <feat name="base"     value="universidade"/>
</struct>
….
```

[9] http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
[10] http://visl.sdu.dk

Pos

```
<struct type="pos">
  <feat name="id"    value="pos1"/>
  <feat name="class"    value="art"/>
  <feat name="gender"    value="F"/>
  <feat name="number"    value="S"/>
  <feat name="canon"    value="o"/>
  <feat name="complement"  value="artd"/>
  <feat name="tokenref"    value="t1"/>
</struct>
```

Phrases

```
<struct type="phrase" from="t1" to="t2">
 <feat name="id"     value="phr1"/>
 <feat name="cat"    value="NP"/>
 <feat name="function" value="subj/>
 <feat name="head"   value="t2"/>
</struct>
…
```

Figure 2: Syntactic annotation

3. *Portal de Córpus*: Site Architecture

The *Portal de Corpús*[11], an XCES compliant corpus portal developed within the PLN-BR project, provides a tool suit for basic facilities to store and retrieve text conformant to XCES format. It was designed using open source technologies and can be easily ported to other servers.

The architecture chosen was client-server (see Figure 3), where in one side we have as interface a web site and in the other side a web server and a database.

The server side was implemented using Java programming language, because Java is cross-platform and it is supported by all web servers. In our implementation, we chose the Apache Tomcat[12] web server as it is the most used and stable servlet container. The site was developed using Jsp and Servlets, and we utilized the MySQL[13] database server that is freely available.

In the client side the user can access the site through a web browser. In the site, we made available an interface where users should register to have access to the main functionalities of the *Portal de Córpus*. A user will be able, for example, to use a Header editor to insert new texts into the database or update headers of texts already inserted.

This portal is based on a database that maps the XCES format into relational entities. The texts and the headers information are inserted into the database using a XCES-compliant header editor (Section 4). Our framework supports multiple corpora in one portal, but the information of each text corpus must be saved in a different database. In the tables of the portal database an administrator can define what corpus will be public for the registered users. The interface that will manage these functions is under development.

[11] See http://www.nilc.icmc.usp.br:8180/portal/
[12] See http://tomcat.apache.org/
[13] See http://www.mysql.com/

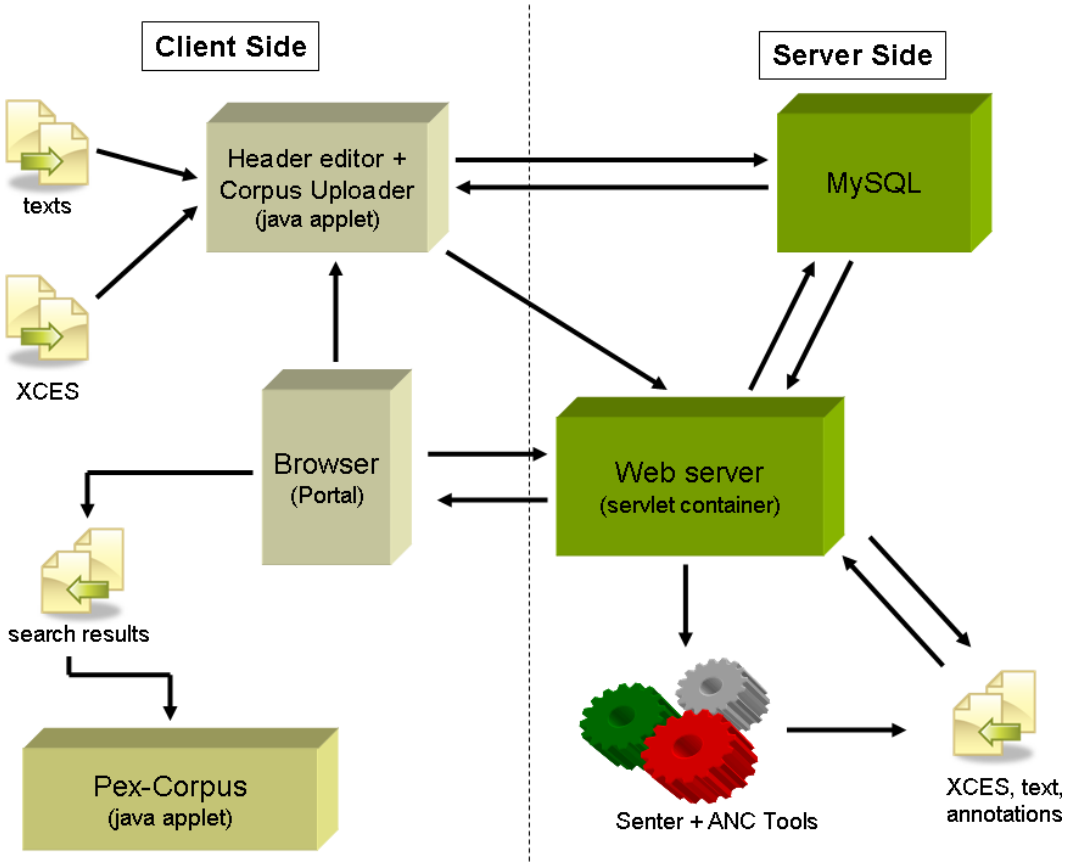

Figure 3: The Portal client/server architecture

After a text is inserted into the database, a special server-side function is called that automatically creates two stand-off annotations (the *Logical markup* and *Sentence boundaries* annotations) for it. In addition, a merged version of the text with these two annotations is created using the program based on the ANC tools library. Our system always keeps one copy of each annotation, of the text and of the text header in both the file system and in the database. This solution was used to optimize the response time for searches in the site since we do not need to process the files for each search.

We provide several search functions to build study corpora from a main corpus. There is an interface for the user to select a corpus to generate a subcorpus of it. The searches are based on the information present in the text headers, such as bibliographic information, newspaper sections[14], text types and keywords. Figure 4 shows a search based on keywords. The user can look for a specific keyword to be used by selecting the link named "keywords list" on the right-hand side of the keyword text box. To limit the search scope, the user can choose the publication year before recovering the matching texts. The searches can return results in several formats, and the user has the option to select one of them. The possible result formats are shown in Figure 4: only the text; the text, the header and all the stand-off annotations; and the merged text with the annotations *Logical markup* and *Sentence boundaries*. All results are compressed and returned in the zip format.

While the major part of the information contained in the header is based on external text information, topic is one of those that should be recovered based on internal information. Therefore after generating a subcorpus the user can use the *PEx-*

[14] This option was particularly designed for newspaper corpora.

Corpus Tool, where she/he can visually inspect the subcorpus already created to explore its content and create further subcorpora based on a selection of text topics (see further information on Section 5).

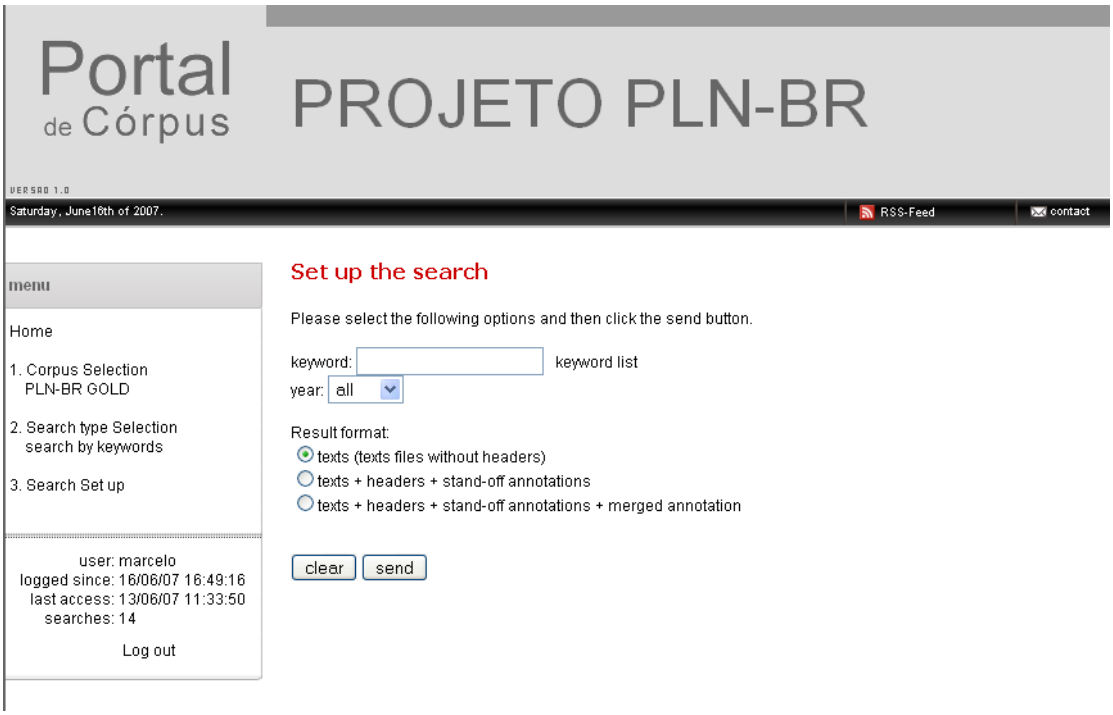

Figure 4: Search by keyword set up

We also made available in the site instructions to install the *Portal de Córpus* in other servers[15]. The source code, the database structure and documentation are freely available.

4. Header Editor and Corpus Uploader

The header editor is a tool implemented using Java applets. It has a graphical interface that allows the user to create, maintain and visualize text header information that is stored in a MySQL database. To access a given database, the structure of the corpus database must follow the structure specified in this project, including the Text Typology used, which in turn follows the one used in the Lácio-Web project (Aluísio et. al, 2003).

Taking into account that the *Portal de Córpus* can store several corpora, and that each corpus has a separated database, the first step to use the editor is to establish a connection with the specific corpus database. The user must know in advance database connection information such as: host URL or IP, user and password, and the database name to access it (Figure 5).

[15] http://www.nilc.icmc.usp.br:8180/portal/downloads.jsp

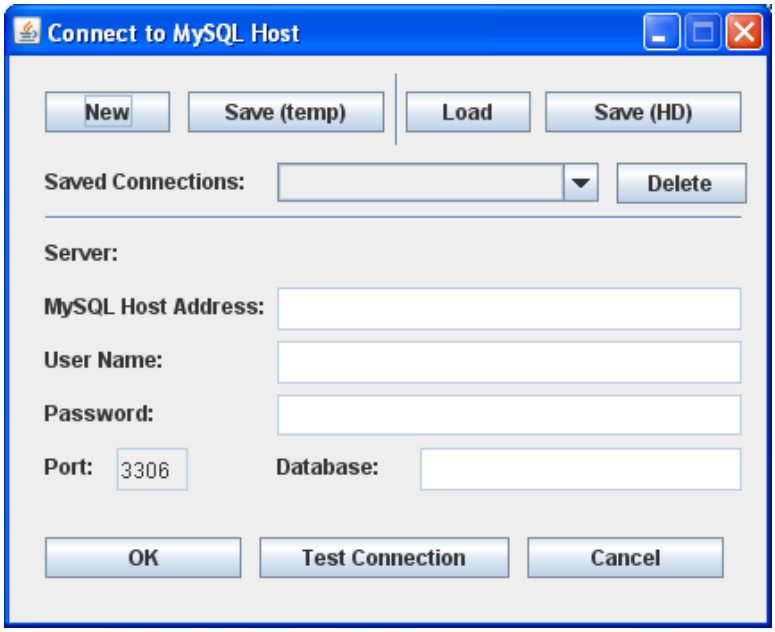

Figure 5: Database connection set up

After connection to the database, the user has three options (Figure 6) to open a text: open a text file, open a text already saved into the database or open a valid XCES XML document.

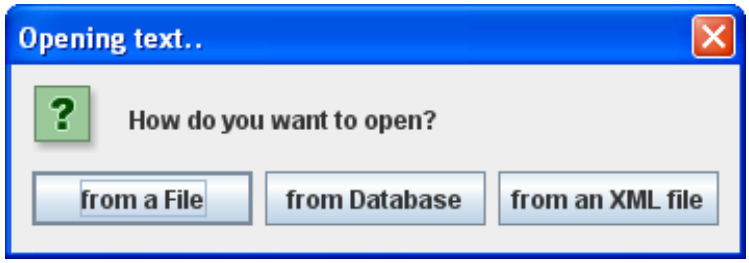

Figure 6: Options to open a text.

Right after the user selects one option to open a text, the header editor will load the file and the user will have the option to edit the header entering all the available information about that text.

The option to open a text from an XML file will parse the XML file that must be in XCES format. It will automatically pre-enter the header field values in the header editor based on the information found in the XML file. The user will be able to enter more information or edit the already pre-entered data. If the file is not in the XCES format the operation will fail.

The header editor uses the same text classification as the Lácio-Web project that is a four-category typology where the texts can be classified by genre, textual type, domain and medium of distribution (Figure 7).

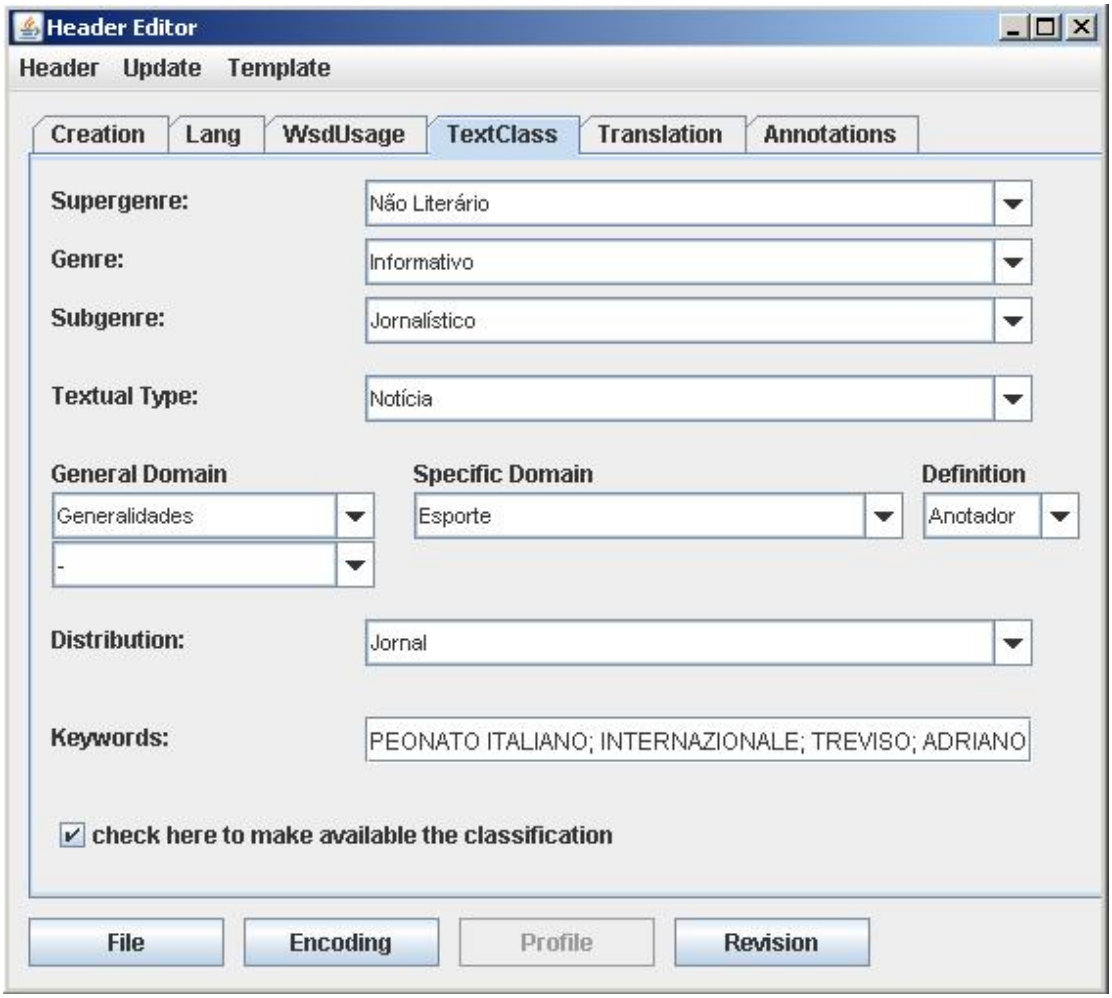

Figure 7: Text classification interface

If the user selects the option to open a text or open a text from an XML file, after editing the header, he will have the option to send the text and the header information to the database (Figure 8). This option will insert all data he/she entered, including the text, into the database he/she selected previously. After inserting a text, the user will always have the option to load the text and its header to update its information.

Figure 8: Sending the header data to the database

The option to open a text from the database allows the user to edit texts that were previously sent to the database. These texts and their header information can be edited and saved using the option update.

The header editor also has an option to insert several texts at once, it's called "upload directory". The requirement to use this functionality is that all header files (extension xces.xml), texts and annotations must be in the same folder. Then the user

must inform the folder pathname and the application will automatically detect the header files, parse them and insert the data in the database (Figure 9).

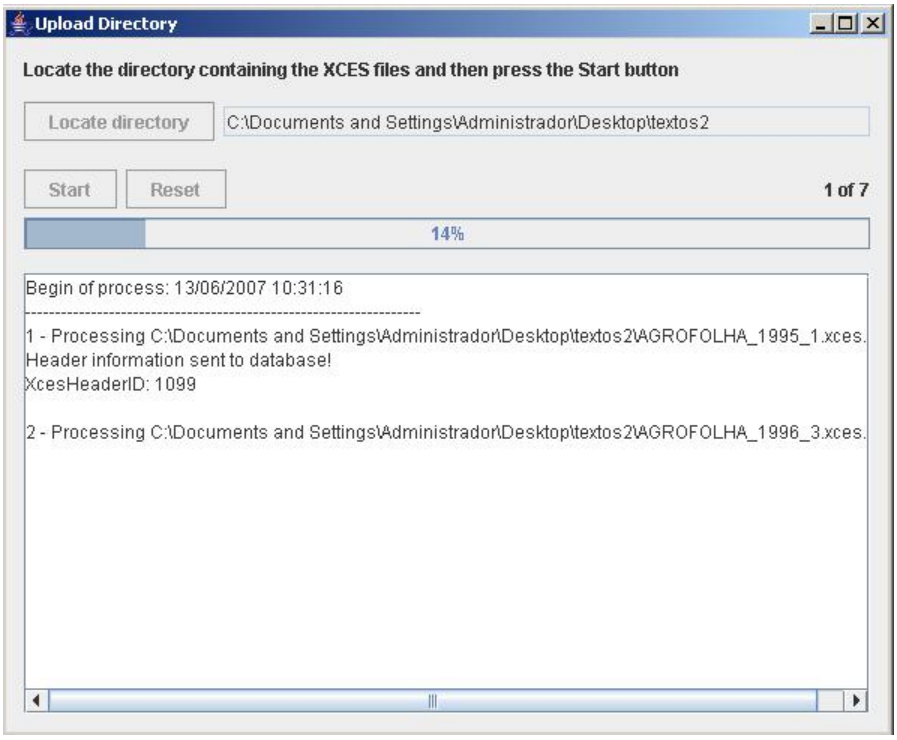

Figure 9: Inserting several texts at once to the database

To speed up the upload process, there is an option that allows the user to create or load header templates. The template is an xml file that can have all the data for one header, so if the user loads it, it will pre-enter those fields inside the template to the header.

5. *PEx-Corpus Tool*

Document collections have increased substantially both in size and complexity, thus extracting useful information from them has become a challenging task. In the pursuit of solutions capable of helping users to explore large document sets, the document map approach is gaining strength.

A document map is a visual representation for user navigation that, similarly to geographical maps, spatially reflects one or multiple properties of the documents that may be of interest. A document map may be built from extracted information, such as co-citations, common citations, co-authorship, and so on (Borner, 2003). When this kind of information is not provided with the documents, content similarity can be employed. In this case, the most promising techniques to create a document map are the so called multidimensional projection techniques.

A multidimensional projection technique typically maps data from a multidimensional space (a space with more than 3 dimensions) into a 1D, 2D or 3D space, whilst retaining, on the projected space, some information about distance relationships among the data items in their original definition space. In this way, a

graphical representation can be created to take advantage of the human visual ability to recognize structures or patterns based on similarity, such as clusters of elements.

Formally, let $X = \{x_1, x_2, ..., x_n\}$ be a set of multidimensional data, with $\delta(x_i, x_j)$ a dissimilarity (distance) measure between two multidimensional data instances, and let $Y = \{y_1, y_2, ..., y_n\}$ be a set of points into a *p*-dimensional space, with $p = \{1,2,3\}$ and $d(y_i, y_j)$ a (Euclidean) distance between two points of the projected space. A multidimensional projection technique can be described as a bijective function $f : X \rightarrow Y$ that seeks to make $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ as close to zero as possible, $\forall x_i, x_j \in X$.

In order to be able to apply a multidimensional projection technique to a document collection it is necessary to determine a way to measure the dissimilarities amongst documents (the above δ). *PEx-Corpus Tool* employs the vector space model (Salton, 1991) to represent the documents as vectors in a multidimensional space, and a cosine-based distance, defined in (Faloutsos, 1995), to determine the dissimilarities amongst the documents as the distance between the vectors that represent them. In the vector space model, the terms that occur in the collection are the space dimensions, and the frequencies of these terms in each document are the coordinates. The process used on *PEx-Corpus Tool* involves three main steps: (i) removing stopwords, i.e., non-informative words such as articles, prepositions and such, plus any words known to lack relevance to context (the stopword list can be defined by the user); (ii) frequency counting, so as to remove terms that occur too sparsely or too often and hence have little differential capability (Luhn's cut-off (Luhn, 1968)); and (iii) weighting the terms according to the term-frequency-inverse-document-frequency (tfidf) measure (Salton, 1991).

On *PEx-Corpus Tool*, the parameters employed to create the vector space model can be defined by the user through a window where it is possible to change the number of grams (one-, bi- or tri-gram) and the stopword list, checking the results on the Zipf's curve of the document collection. Also, it is possible to change the Luhn's upper and lower cut-off and verify the resulting terms and the part of the curve that these cut-offs comprise. Figure 10 shows this window. The rectangle over the curve indicates the part used according to the cut-offs.

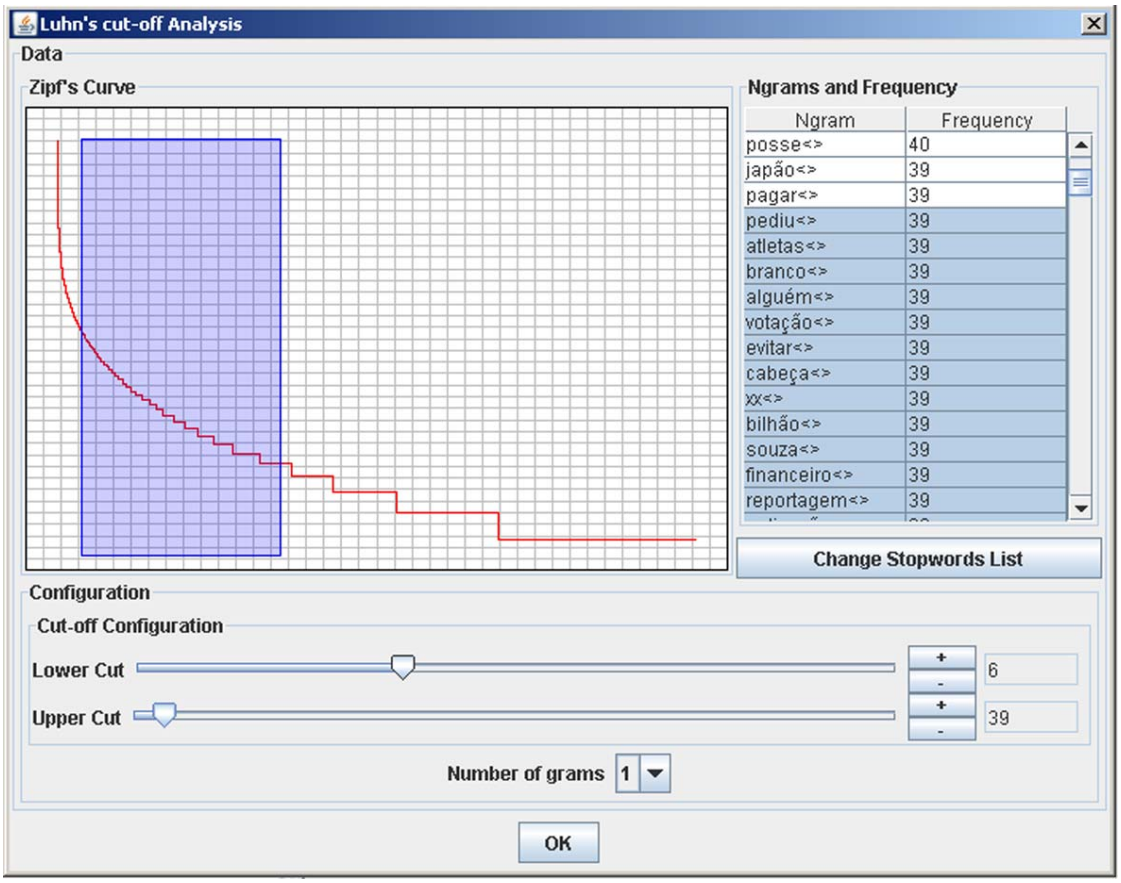

Figure 10: Window to set the parameters of the vector space model.

There exist a number of projection techniques which can be applied on multidimensional data. Normally, a high precision technique, that is, a technique which can preserve as much as possible the distances of the multidimensional space into the projected space, is computationally expensive. On the other hand, less precision techniques can result on poor layouts due to the approximations employed in order to reduce the complexity, specially when working with multidimensional spaces with a high number of dimensions such as the one created by the vector space model. Here we employ a technique, called ProjClus (Paulovich, 2006), which tries to make a trade-off between computational complexity and the resulting layout quality. ProjClus was developed as a part of a document collection visualization tool, called Text Map Explorer (Paulovich, 2006), and has shown to be effective on creating useful document maps.

The result of the projection process is a graphic representation where each document is identified as a point on a plane. Points closed placed indicates documents with similar content, and points far projected indicates non-correlated documents. In order to help users identifying document similarity, document neighborhood information is graphically represented as edges between points. On *PEx-Corpus Tool*, a user may choose to visually connect points on the map either with their nearest neighbors on the 2D plane, or with their nearest neighbors on their original multidimensional space, as computed from the document vector representation. Figure 11 presents the *PEx-Corpus Tool* main window with a document map.

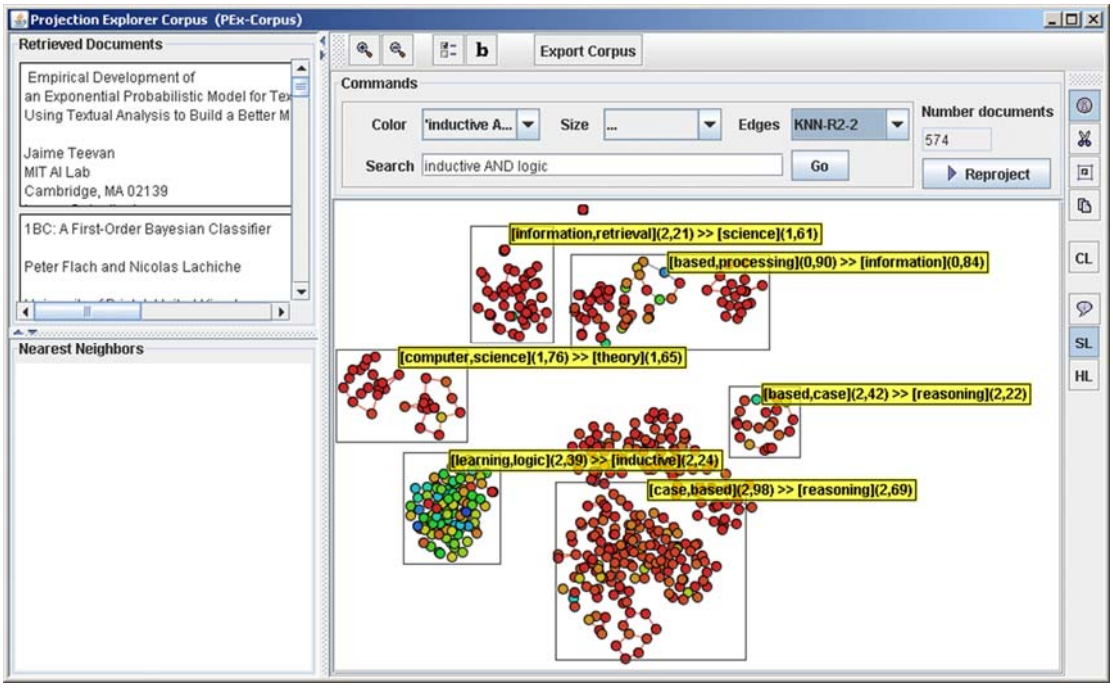

Figure 11: *PEx-Corpus Tool* main window.

PEx-Corpus Tool offers several features in order to explore the document map. If a user rolls the mouse over one point, a label identifying its corresponding document is shown. A single mouse click over a point gives the user a list of its nearest neighbors (either on the plane or on the original multidimensional space, depending on the user choice above), and a double click returns a display of the contents of the document and its neighbors to support further in-depth analysis.

Besides the information on document similarity intrinsically conveyed by the projection, the frequency of occurrence of a word or group of words (chosen by the user) in the documents can be mapped to the color or size of points. Figure 11 is colored based on the frequency of "inductive AND logic".

In order to enable visual identification of the main topics discussed in the documents of the collection, a projection area can be selected - delimiting a region with the mouse - and a label is generated that is representative of the documents within this area. The label is formed by three different terms that occur on the selected documents. The first step to generate the label is to create a document vector representation of the selected documents. Then, the first two label terms are chosen as those with the highest covariance in the vector representation. The third term is the one that has the highest covariance with respect to the two terms already chosen. One can also create clusters over the projection and associate labels to them. On Figure 11 the boxes represent the labels and the rectangles indicate the documents used to create them.

Using the above features, the user can navigate the map discovering the main topics discussed on the document collection, identifying and selecting the most relevant documents according to a certain subject. This enables a step-wise refinement process, where the user starts with a large collection of documents and finishes with a small portion that corresponds to a user claim. Figure 12 outlines this refinement process. On the step (1) the vector space model parameters are chosen. After that a document map is generated on step (2) and the user can interact with on step (3), coloring/sizing the points and creating labels in order to identify the main topics handled on the collection. Based on the insights acquired with this interaction, the

user can cut-off irrelevant documents on step (4), and export the resulting documents as a new document collection on step (5), or re-start the process using only the remaining documents on step (6).

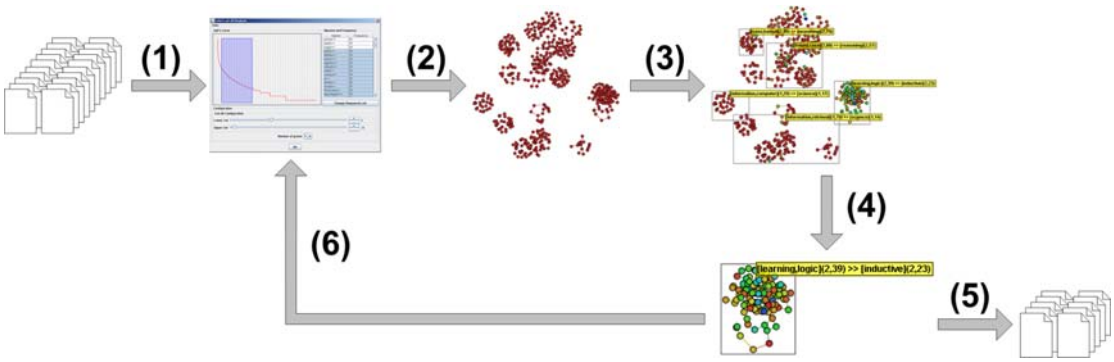

Figure 12: A step-wise refinement process which can be applied with *PEx-Corpus Tool* in order to create a small document collection starting from a bigger one.

6. Conclusion and Future work

In this paper, we presented the *Portal de Córpus*, an XCES compliant corpus portal, which gives access to several Brazilian Portuguese newspaper corpora compiled in the scope of PLN-BR project. One of them, the PLN-BR GOLD, is freely available in the Web. With regards to the available linguistic annotation of this corpus, Palavras syntactic annotation has been used for the study of linguistic informed text classification experiments (Gonçalves et al, 2006) as well as for coreference and anaphora resolution studies and experiments (Coelho et al, 2006; Vieira et al 2005). While the first has made use mainly of part-of-speech data, the second involved information of syntactic structure. As this kind of annotation has been proved useful for these other linguistic level research we consider that providing a release of the GOLD corpus with this annotation following the XCES standard will be useful for the community interested in Portuguese processing. As for the *Portal de Córpus*, one of the main features of it is a tool to allow a user to visually inspect a corpus to explore its content and create study subcorpora based on a selection of topics. This tool, named *PEx-Corpus Tool*, is activated after a user has created a subcorpus based on several information presented in the texts headers. The whole framework can be easily used in other projects of BP corpora using XCES standard, since we also made available a header editor and corpora uploader to import and edit full XCES-compliant headers. In the near future, we intend to make several basic corpus tools available to users. Philologic[16] can be an easy and inexpensive option to work with the in-line annotation version of our corpora as it has concordancers, a word frequency counting tool, and other tools such as similarity search and collocation reporting.

[16] http://philologic.uchicago.edu/index.php

References

Aires, R., A. Mamfrin, S. Aluísio and D. Santos. (2004) 'What is my Style? Using Stylistic Features of Portuguese Web Texts to classify Web pages according to Users' Needs.' *In LREC 2004*. pp. 1943–46. Lisbon, Portugal.

Aires, R., S. Aluísio and D. Santos. (2005) 'User-aware page classification in a search engine.' *In SIGIR Workshop on Textual Stylistics in Information Access*. Salvador - Brazil.

Aluísio, S. M. G. Pinheiro, M. Finger, M. G. V. Nunes and S.E. Tagnin. (2003) 'The Lacio-Web Project: overview and issues in Brazilian Portuguese corpora creation.' *In: Corpus Linguistics 2003: Proceedings of Corpus Linguistics 2003 (Also as UCREL Technical Report, Vol 16 Part)*. Lancaster: 2003. v. 16, pp. 14–21.

Biber, D. (1993). 'Representativeness in corpus design'. Literary and Linguistic Computing 8, pp. 243–57.

Bick, E. (2002) The Parsing System "Palavras" - Automatic Grammatical Analysis in a Constraint Grammar Framework. Århus: Aarhus University Press

Borner, K., C. Chen and K. Boyack. (2003) 'Visualizing knowledge domains'. *Annual Review of Information Science and Technology*, 37: 1–51.

Burnard, L. (1995) An Introduction to the Text Encoding Initiative. TEI Document no TEI J31, Oxford University Computing Services. Available at http://www.tei-c.org/Vault/SC/J31/ (accessed: 28 June 2007)

Coelho, J. B., V. M. Muller, S. C. de Abreu, R.Vieira and L. H. M. Rino. (2006) 'Resolving Nominal Anaphora'. In: *7th Workshop on the Computational Treatment of Portuguese Language*, Itatiaia. Lecture Notes in Artificial Intelligence. Berlin: Springer. v. 3960. pp. 160–69.

Eagles - Expert Advisory Group on Language Engineering Standards (1996) *Preliminary recommendations on text typology*. EAGLES Document EAG-TCWG-TTYP/P. Available at http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html (accessed: 28 June 2007)

Faloutsos C. and K. Lin. (1995) 'Fastmap: A fast algorithm for indexing, data mining and visualization of traditional and multimedia databases'. *In* Proceedings of the 1995 ACM SIGMOD international conference on Management of data, pp. 163–74, San Jose-CA, USA. ACM Press: New York.

Gasperin C., R. Vieira, R. Goulart and P. Quaresma. (2003) Extracting XML chunks from Portuguese Corpora In: *Proceedings of the Workshop on Traitement automatique des langues minoritaires*. Batz-sur-Mer: ATALA, v.2. pp. 223–32.

Gonçalves, T., C. Silva, P. Quaresma and R. Vieira. (2006) Analysing Part-of-Speech for Portuguese Text Classification . In: *Computational Linguistics and Intelligent Text Processing: 7th International Conference*, CICLing, Mexico City. Lecture Notes in Computer Science. Berlin : Springer. v. 3878. pp. 551–62.

Ide, N. and C. Brew. (2000). 'Requirements, Tools, and Architectures for Annotated Corpora.' *Proceedings of Data Architectures and Software Support for Large Corpora.* Paris: European Language Resources Association, 1–5.

Ide, N. and L. Romary. (2004). 'International standard for a linguistic annotation framework.' *Journal of Natural La Language Engineering*, 10: 3–4, 211–25.

Ide, N., P. Bonhomme and L. Romary. (2000). ‘XCES: An XML-based Standard for Linguistic Corpora.’ *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, pp. 825–30. Athens, Greece.

Luhn H.P. (1968) ‘The automatic creation of literature abstracts’. *IBM Journal of Research and Development*, 2(2): 159–65.

Paulovich F. V. and R. Minghim. (2006) ‘Text map explorer: a tool to create and explore document maps’. *In IV ‘06: Proceedings of the conference on Information Visualization*, pp. 245–51, Washington, DC, USA. IEEE Computer Society.

Salton G. (1991) ‘Developments in automatic text retrieval’. *Science*, 253:974–80.

Vieira R., C. Gasperin, R. Goulart and S. Salmon-Alt. (2003) ’From concrete to virtual annotation markup language: the case of COMMOn-REFs’ In: *Prooceedings of ACL Workshop on Linguistic Annotation: Getting the Model Right***,**. Sapporo: ACL, pp. 6–13.

Vieira R., S. Salmon-Alt and C. V. Gasperin. (2005) ‘Coreference and Anaphoric Relations of Demonstrative Noun Phrases in Multilingual Corpus’. In: *Anaphora Processing:linguistic, cognitive and computational modeling***.** Ed.Amsterdam : John Benjamins. pp. 385–403.