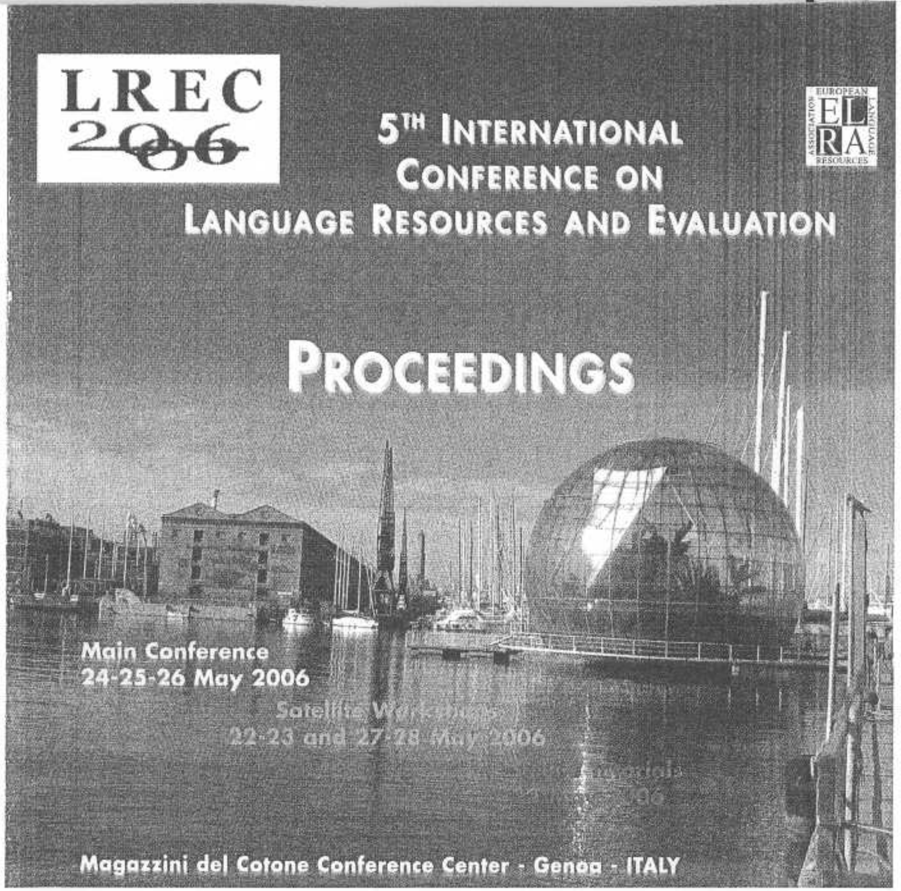
LREC 2006 is hosted by
Istituto du Linguistica Computazionale del Consiglio Nazionale delle Ricerche (ILC-CNR)
Head Office in Pisa - Branch Office in Genoa - ITALY

European Language Resources Association (ELRA)
55-57, rue Brillat-Savarin - 75013 Paris - FRANCE
Tel.: +33 1 43 13 33 33 - Fax: +33 1 43 13 33 30 - Web: http://www.elra.info - http://www.elda.org

ISBN 2-9517408-2-4
EAN 9782951740822

Edited by ELRA - Graphic design by Paola Baroni (ILC-CNR, Pisa)
Front cover photo by Lucia Marconi (ILC-CNR, Genoa) - Back cover photo by Eugenio Mirici (Genoa)



LREC 2006

5TH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

PROCEEDINGS

Main Conference
24-25-26 May 2006

Satellite Workshops
22-23 and 27-28 May 2006

Tutorials
[illegible]

Magazzini del Cotone Conference Center - Genoa - ITALY

Retrieving Terminological Data from the TxtCeram Tagged Domain Corpus: A First Step towards a Terminological Ontology

Anna Estellés, Amparo Alcina, Victoria Soler

Department of Translation and Communication
Faculty of Human and Social Sciences
Universitat Jaume I - Campus de Riu Sec
E-12080 Castellón de la Plana
Spain
anna.estelles@alumail.uji.es, alcina@ trad.uji.es, victoria.soler@alumail.uji.es

Abstract

In this paper we will focus on corpora as a resource for researching language processing for terminological purposes. Based on the TEI guide, we present the templates used to tag our TxtCeram corpus and its features when working with WordSmith, a text analysis tool. We present an experiment for studying the frequency of hyperonyms in the introduction section of texts, while testing WordSmith's suitability to work with our tagged corpus.

1. Introduction

The use of corpora has become a valuable resource in areas dealing with natural language processing, including the use of corpora by language practitioners as focus of their studies or terminological research. In this paper we will focus on corpora as a resource for researching language processing for terminological purposes. In terminology, researchers explore domain-specific language material to investigate terminological issues and thus, corpora represent a realistic model of language to be studied (Bowker, 1996). That is, the need for linguistic research based on the use of corpora is due to a need of studying language through real examples (Sinclair, 1991). Tognini-Bonelli (1996) describes the diversity of applications of corpus linguistics and offers a wide explanation about uses and targets of corpus linguistics. The author defends the use of corpora as a basis for an empirical model in research, consisting on the observation of language facts, the formulation of hypotheses and generalizations based on patterns of data, and the subsequent derivation of theoretical statements (Tognini-Bonnelli, 2001). Another aspect to be considered is the treatment that corpora have received. The existence of tools to analyse text, parsers, as well as codes or standards to encode texts in order to make texts machine-readable have empowered the scope of corpora-based studies (see Garside, Leech & McEnery, 1997). Authors such as Ahmad and Rogers (1997) explain that corpora treatment will depend on explicit criteria. Basically, raw corpora can be used to find the frequency of occurrence of single words or multiword compounds, among others. On the other hand, a tagged corpus is a linguistically analysed or annotated version of a raw corpus that may contain not only the constituent words but also part-of-speech information (Ahmad & Rogers, 1997). Moreover, among others, tagging can target semantic, morpho-syntactic, contextual and/or structural aspects, therefore the aim of the compilation will determine the information that will be tagged (Alcina, 2001). Some computing languages allow text tagging and machine reading; TEI and Docbook are two examples. At present, the TEI (Text Encoding Initiative) provides the TEI Guidelines, which are an international and interdisciplinary standard that facilitates libraries, museums, publishers, and individual scholars represent a variety of literary and linguistic texts for online research, teaching, and preservation (Burnard & Sperberg-McQueen, 2002). The first editions of the Guidelines used the *Standard Generalized Markup Language* (SGML); the last stable edition (TEI P4 from 2002) can also be expressed in the Extensible Markup Language (XML). The TEI standard comes from a part of the Corpus Encoding Standard (CES) proposed by the EAGLES group ("Expert Advisory Group on Language Engineering Standards"). Its simpler collection is TEI Lite, an extensive set of elements and recommendations already based on TEI P4 but simpler in its syntax. DocBook also provides a system for writing structured documents using SGML or XML. Similarly to the TEI, there are different versions of DocBook based on SGML or XML which main structures correspond to the general notion of what constitutes a book (Walsh & Muellner, 1999).

In order to analyze this tagged corpora, computing linguistics and AI research are developing a wide range of tools and resources that can help terminological research and terminography. On the one hand, processes such as the manual acquisition of terminological data from text material are being replaced for machine techniques that help in this work-intensive task. On the other hand, knowledge bases and ontology systems are becoming powerful tools in order to manage terminological data and knowledge (Miller & Fellbaum, 1991; Guarino, 1995; Gruber, 1993).

In this paper we present the design of the TxtCeram corpus and the tagging model adopted. Corpus tagging will be viewed as the first step in order to retrieve data for building a terminological ontology, and as an object of analysis and research for methodological practices aiming to build a terminological ontology Finally, we present some strategies and methodologies order to improve the processing of tagged corpora when using text-analysis tools such as WordSmith.

2. TxtCeram Project

TxtCeram is a research project that is being developed by the Tecnolettra research group[1]. The project focuses on semiautomatic extraction and formal conceptual analyses of ceramic terms. The main objective of the project is to check the efficiency of some of the computing tools used to design an integral system of assisted terminology and the benefits of using this system in linguistic mediation. The project involves testing ontology editors and studying their application in the generation of knowledge bases.

For this purpose, an electronic corpus of specialised texts from the field of ceramics has been compiled. As referred above, the corpus is aimed to terminology building and its target users are linguistic mediators, and especially translators. At present, we have a compiled a corpus of 12,6 MB (txt format) in Spanish and we have begun to compile an English corpus, already being of 1,16 MB (txt format); that is 2,8 million words in the Spanish corpus and about 250000 words in the English one that has recently started to be compiled. A Catalan corpus has begun to be compiled as well. Texts compiled are original works belonging to the field of ceramics. The compilation process has followed an exhaustive criteria of documentation and evaluation based on qualitative research procedures (Alcina, 2005). We have selected and digitalized books and works that, from our terminological point of view, represent a very important knowledge in terms of communicative situations. Moreover, we have focused in collecting those books that may contain not only relevant terminology but also relevant contexts and explanations dealing with the processes and the elements involved in ceramics. That leads to a wide range of specialized knowledge that will serve as a base for our research on ceramic conceptual classification, description of concepts, efficacy of tools, etc. The selection included eleven books dealing with ceramics and including data about: products and technologies for glaze and colour materials, manufacturing, uses of tiles, structures, chemicals, materials, procedures and processes.

With regard to its compiling criteria, our corpus currently has the following features: it is written, tagged, synchronic (works have been published during the same period, 1980-1997), specialized, textual (complete works), documented (as it includes not only the text but also part-of-speech information) and multilingual (Spanish, English and Catalan)[2].

3. HTML, TEI P4 Lite, CES & DocBook

Focusing on tag elements, several branches can be distinguished depending on the aim of the tagging process. Despite of the use or not of machine taggers, tagging can be mainly targeted to: morphology, syntax, lexical aspects, semantics and/or discursive aspects. In our work, we explore the tagging of the macro-structure and contextual information of the corpus. As explained above, the corpus will be used as a resource to extract terminological information. These data will serve as the basis for our research tests, including the design of a terminological ontology. The macrostructure of the texts provides information about the textual genre and the parts that make up the document. It can be helpful in the extraction of terminological data, since the frequency of terms and their position may be affected depending on the focus of our research. Contextual information will be helpful in order to understand the characteristics of the text, and thus the environment of the term.

As explained above, TEI Lite and DocBook are two languages for marking up texts. They are very similar since the latest versions of both are based on XML, which implies a DTD (document type description) that can be more or less complex depending on users' needs. On the other hand, there is HTML, the standard language for structuring information in Internet. When analyzing the special features of those languages, it is noticed that although HTML was thought as a way of combining and creating digital resources, its architecture does not support semantic information. DocBook is understood as a DTD to publish works, its architecture is focused on allowing both, electronic and printed representation, so semantic information is designed according to that view. CES and TEI accomplish the same standard guidelines; the main difference can be found in the name of some tags. The TEI was created in order to represent documents that already existed (Rahtz, Walsh & Burnard, 2004), and the CES (Corpus Encoding Standard) is part of the EAGLES Guidelines. Both two are XML-based and have influenced the tagging of the British National Corpus. They are designed to be optimally suited for being used in language engineering research and its applications, in order to mark up structural information, and to add semantic information to texts. The CES extends some branches to provide a deeper level of description (such as sentence tags and keywords in sentence); while the TEI presents a more superficial analysis (structure and communicative situation). Nevertheless, when dealing with structural and main semantical information both share the same tags. In our work we have drawn on TEI.4, which differences with the CES remain the same. We have chosen that standard because the information we will mark is already implemented in both languages, and it is simplified and fully documented in the TEI. We do not decline the use of the CES in future works if the detail of the mark-up requires this other standard.

4. Tagging

In order to systematize the process of tagging we have elaborated two templates: one for the books, and another for future articles we will add to the corpus. The templates can be divided in two parts: the former marks the contextual information and it is the same for both, books and articles; the later refers to the macrostructure and it has differences, as it will be observed.

4.1. Contextual Information

As it is showed in Table 1, contextual information is distributed in six groups where tags are included, those groups will be tagged inside the <head> of the xml document and are the following ones:

[1] Project developed by the Tecnolettra research group, Translation and Communication Dept., Jaume I University. http://tecnolettra.uji.es.

[2] For more information about corpus typology see Sinclair, 1991; Tonigni-Bonelli, 1996 & 2001; McEnery & Wilson, 1996; Garside, Leech & McEnery, 1997; Biber, Conrad & Reppen, 1998; Reppen, Fitzmaurice & Biber, 2002; or Sánchez-Gijón, 2004; among others.

Information about the file – It contains a file description, including the format, the extension and identifying number (idno).

Information about the creation of the electronic document – It contains the creator of the document, the publisher and distributor of the digital work, the place where it has been created and the date. This information obviously differs from the one of the original document.

Information about the original document – Here, all the bibliographic data of the original work is included.

Information about the tagging – It contains the encoding description and the encoding project description.

Information about the type of document – This section includes information about the communicative situation of the original work. In addition, the language of the text that will serve to distinguish the documents is marked here. The subject, keywords, a class code based on the Library of Congress Classification, participants and mode are also included in this section.

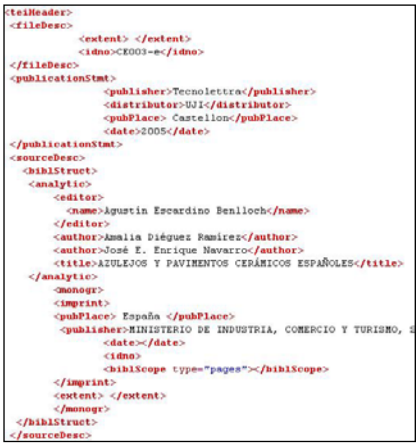
Information about the reviewing process of the documents – In order to manage easily the work of compiling and tagging the text, this section includes information about changes made, responsible of changes, etc.

	Tagging	Function	Compulsory	Path
File information	<fileDesc>	File description	Yes	Book or article, <head>
	<extent>	File size in Kb	No	Book or article, <head> , <fileDesc>
	<idno>	Code for identifying the file	Yes	Book or article, <head> , <fileDesc>
Electronic file description	<publicationStmt>	Information about the creator of the electronic document	Yes	Book or article, <head>
	<publisher>	Information about the publisher of the electronic document	Yes	Book or article, <head>, <publicationStmt>
	<distributor>	Information about the distributor's name of the electronic document	Yes	Book or article, <head>, <publicationStmt>
	<pubPlace>	Place where the publication has been developed	Yes	Book or article, <head>, <publicationStmt>
	<date>	Date when the electronic document was created	Yes	Book or article, <head>, <publicationStmt>
Original work information	<sourceDesc>	Bibliographical information about the original work	Yes	Book or article, <head>
	<biblStruc>	Structured bibliographic information	Yes	Book or article, <head>, <sourceDesc>
	<analytic>	Information about the text included in the file (it refers to the specific part, i.e. if it is an article, a chapter of a book or a full book)	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>
	<author>	Text's author if different from the work it belongs to	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <analytic>
	<title>	Title of the original document	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <analytic>
	<monogr>	Information about the original work to which the document belongs (book or journal)	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>,
	<editor>	Editor/s of the work, it might contain a <name> tag for names	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>
	<imprint>	Original work press information	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>
	<pubPlace>	Place where the original work was published	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>, <imprint>
	<publisher>	Information about the publisher	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>, <imprint>
	<date>	Date of publication of the original work	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>, <imprint>
	<idno>	ISBN or ISSN of the original work	Yes	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>, <imprint>
	<biblScope type="pages">	Number of pages of the original text	No	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>, <imprint>
	<title>	Title of the original work. It is compulsory if it is different from the one of the tagged document (articles or chapters)	-	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>
	<author>	The original work author/s. It is compulsory if it is different from the one/s of the tagged document (articles or chapters)	-	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>
	<extent>	Number of pages of the original work if it is different from the electronic document (articles or chapters)	No	Book or article, <head>, <sourceDesc>, <biblStruc>, <monogr>
Encoding information	<encodingDesc>	Encoding information (for instance xml-TEI)	Yes	Book or article, <head>
	<projectDesc>	Description of the encoding project	Yes	Book or article, <head>, <encodingDesc>
	<editorialDecl>	Legal aspects regarding to the use of this document	Yes	Book or article, <head>, <encodingDesc>
File type information	<profileDesc>	Information about the type of document. Its original communicative situation	Yes	Book or article, <head>
	<creation>	Format of the document: paper, sound, etc.	Yes	Book or article, <head>, <profileDesc>
	<langUsage>	Information about the language used	No	Book or article, <head>, <profileDesc>
	<language>	Way of distinguishing the language/s used in the document	Yes	Book or article, <head>, <profileDesc>, <langUsage>
	<textClass>	Information about the subject of the text based on LCC encoding	Yes	Book or article, <head>, <profileDesc>
	<keywords>	Keywords in order to identify the text. Words might be listed using <list> and <item> tags	Yes	Book or article, <head>, <profileDesc>
	<classCode>	Code that corresponds to the <textClass> description	Yes	Book or article, <head>, <profileDesc>
	<catRef>	Encoding system used for <classCode> and <textClass> tags. LCC for all files	Yes	Book or article, <head>, <profileDesc>
	<textDesc>	Concise description of the text	Yes	Book or article, <head>, <profileDesc>
	<particDesc>	Description of issuer and recipient profiles	Yes	Book or article, <head>, <profileDesc>
	<settingDesc>	Mode of the document: written, oral, etc.	Yes	Book or article, <head>, <profileDesc>
Information about reviews	<revisionDesc>	Events happened when tagging and reviewing the tagging of the text	Yes	Book or article, <head>
	<change>	Description of changes made	No	Book or article, <head>, <revisionDesc>
	<date>	Date when changes were made	No	Book or article, <head>, <revisionDesc>
	<respStmt>	Specification of who made the change	Yes	Book or article, <head>, <revisionDesc>

	<item>	Copy of the modified segment and the original segment	Yes	Book or article, <head>, <revisionDesc>

Table1. Contextual information

A sample of how header looks like after fulfilling it is presented in Figure 1.



```
<teiHeader>
 <fileDesc>
            <extent> </extent>
            <idno>CE003-e</idno>
 </fileDesc>
 <publicationStmt>
                <publisher>Tecnolettra</publisher>
                <distributor>UJI</distributor>
                <pubPlace> Castellon</pubPlace>
                <date>2005</date>
 </publicationStmt>
 <sourceDesc>
   <biblStruct>
    <analytic>
        <editor>
          <name>Agustin Escardino Benlloch</name>
        </editor>
        <author>Amalia Diéguez Ramirez</author>
        <author>José E. Enrique Navarro</author>
        <title>AZULEJOS Y PAVIMENTOS CERÁMICOS ESPAÑOLES</title>
    </analytic>
        <monogr>
        <imprint>
        <pubPlace> España </pubPlace>
         <publisher>MINISTERIO DE INDUSTRIA, COMERCIO Y TURISMO, S
                <date></date>
                <idno>
                <biblScope type="pages"></biblScope>
        </imprint>
        <extent> </extent>
        </monogr>
  </biblStruct>
 </sourceDesc>
```

Figure 1. Header segment

4.2. Macrostructure

With regard to the macrostructure, we have simplified the collection of tags in order to achieve two schemes. This description is based, mainly, on the theories based on textual genre (see Swales, 1990; Bahtia, 2002). Those authors defend the importance of the purposes of the community of speakers of a specialized language, suggesting that a specialized language is characterised by a set of communicative purposes agreed by the members of the discursive community.

Macrostructure elements are marked inside a <text> tag and differ depending on the textual genre of the original work. The variations at the <div> tag attribute description in the DTD allow the representation of the variations in the description of a book compared to an article. Table 2 includes the tags that mark an article and Table 3 the ones that mark a book. If we compare them, we will notice that the strategy that has been held is to allow a parallel working of both structures. The target is to allow inter-textual comparison of sections and give flexibility when working in certain sections of the texts in the corpus. The data included in the tags describe sections and headings.

Encoding	Function	Path
<front>	Preliminary information	<text>
<div type= “abstract”>	Marking abstract section	<text>, <front>
<div type= “index”>	Marking index section	<text>, <front>
<body>	Marking the body of the article	<text>
<div type=”1” >	Marking the introduction of the article	<text>, <body>
<div type=”conclusion” >	Marking the conclusion section	<text>, <body>
<head>	Marking a title (at any level)	<text>, ...
<back>	It contains acknowledgements, appendix section/s and bibliography	<text>
<div type= “appendix”>	Marking appendices or tables added at the end of the document	<text>, <back>
<div type= “bibliography”>	Marking bibliography section	<text> <back>

Table 2. Article

Encoding	Function	Path
<front>	It contains preliminary information: preface, acknowledgments, index...	<text>
<div type=”preface”>	Marking preface section	<text>, <front>
<div type=”index”>	Marking index section	<text>, <front>
<body>	It contains the body of the document	<text>
<div type=”1” >	Marking the introduction section	<text>
<head>	Marking a title (at any level)	<text>, ...
<back>	It contains acknowledgements, appendix section/s and bibliography	<text>
<div type=”appendix”>	Marking appendices or tables added at the end of the document	<text>, <back>
<div type=” bibliography”>	Marking bibliography section	<text> <back>

Table 3. Book

5. Processing

Automatic or semi-automatic term extraction is considered an accelerating factor in the research process of long-term creation of professional resources; as in our case or as it has been done in other works. The use of verbs in scientific articles has been studied using frequency lists to help to identify data, another example is the study of the usefulness of frequency lists in terminographical methodology (Reimerick, 2002; Pérez Hernández, 2002; Faber 2002). In this section, we present some strategies to process tagged information with a text analysis tool in order to obtain benefits from contextual and macrostructural information.

5.1. WordSmith Tool

WordSmith is a textual analysis tool that works with raw text and simple tagged text. It was developed by Mike Scott. This tool is useful in linguistic engineering, an example of its use is its application based on the textual genre concept (Alcina, 2005). WordSmith includes three tools (we are using WordSmith 3 version):

WordList – provides word lists or groups of words from a text in an alphabetical order or in a frequency order. It allows statistical treatment: rate of length of words, sentences or segments; number of words depending on their number of letters; types, tokens and their relation.

Concord – provides word lists and their context.

KeyWords – allows the search of keywords in a text.

These tools allow the use of stop list and lemma files. A stop list is a list of words that must not be taken into account during the calculating processes. Lemma files include lists of words grouped into a same lemma that will be counted as a same entry.

The configuration of this tool in order to benefit from tagged information depends on the objective of the research.

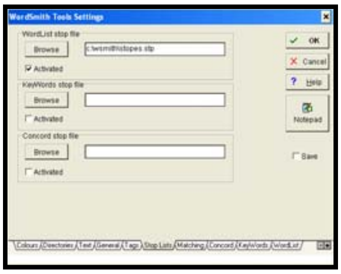
5.2. Our experiment

As a sample, we configured the tool to provide a frequency word list of the introduction section to be compared with a frequency word list of the body section. Our sample is based on 16 files including 4 books. Each chapter of each book was tagged according to templates explained earlier and so was done for the book. As a result, a tagged corpus of 306.068 words including tags was created, using WordList to perform the statistical analysis

Once files were selected, we applied a stop list and a lemma file. The settings used are shown in Figures 2 and 3, respectively. The Txtceram project has compiled a stop list for the Spanish corpus that includes pronouns, articles, prepositions, numerals, relatives and common adverbs. The Txtceram lemma file includes a list of 5548 Spanish verbs. It consists of a compilation of verbs extracted from the collection made by Alonso (Alonso, 1989).

We designed a tag file that included tags that we wanted to be taken into account and ignored tags that must not be considered (Figure 5). This removed noise to the corpora since tags such as contextual ones (<author>, <analytic>, etc.) were not counted as words of the text. Our tag file consisted of the tag for identifying introductions (<div type= "1">) and the one for marking up the body (<body>) and their respective close tags.

Once the tag file was created, we configured the tool to load our tag file (Figure 5). Next step was selecting the section we wanted to use to calculate the frequency list. A first calculation included the introduction section (Figure 6) and then, a second calculation of the body section was made, where we replaced the <div type="1"> and </div> tags with the <body> and </body> tags. Results (Figure 7 and Figure 8) showed that the introduction word list presented hyperonyms at the top positions with a high rate of frequency. When compared to the body section, we distinguished a lack of descriptors in the introduction sections. The context of the words showed that the 20 first words in the list were not followed by descriptors. In contrast, in the body section word list, the context of the first 10 token included descriptors in a near position and from the 11th position we started to view terms that belong to more specific areas, such as *bizcocho* (biscuit), *esmalte* (glaze) or *rodillo* (roller). Provisionally, and though more in-depth research has to be conducted in order to state this, we concluded that the introduction sections of ceramic works are not remarkable for finding descriptors, but are useful to detect hyperonyms.


WordSmith Tools Settings

WordList stop file: c:\wsmith\stopes.stp
☑ Activated

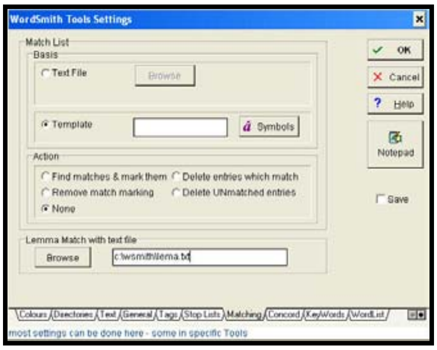
KeyWords stop file:
☐ Activated

Concord stop file:
☐ Activated

☐ Save

Browse
OK
Cancel
Help
Notepad
Colours Directories Text General Tags Stop Lists Matching Concord KeyWords WordList

Figure 2. Stop list


WordSmith Tools Settings

Match List
Basis
○ Text File
● Template

Action
○ Find matches & mark them ○ Delete entries which match
○ Remove match marking ○ Delete UNmatched entries
● None

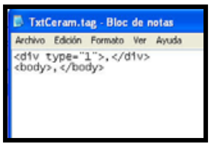
Lemma Match with text file: c:\wsmith\lema.txt

☐ Save

most settings can be done here - some in specific Tools

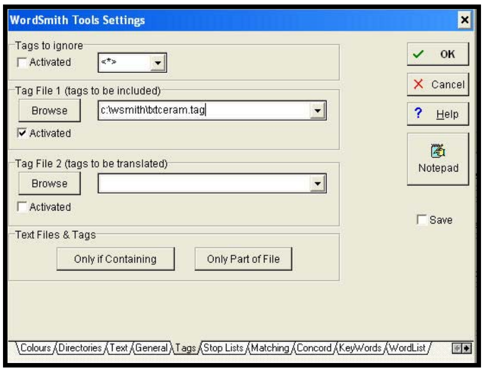
Browse
Symbols
OK
Cancel
Help
Notepad
Colours Directories Text General Tags Stop Lists Matching Concord KeyWords WordList

Figure 3. Lemmatizing


TxtCeram.tag - Bloc de notas

```
<div type="1">,</div>
<body>,</body>
```

Figure 4. Tag list


WordSmith Tools Settings

Tags to ignore: <*>
☐ Activated

Tag File 1 (tags to be included): c:\wsmith\txtceram.tag
☑ Activated

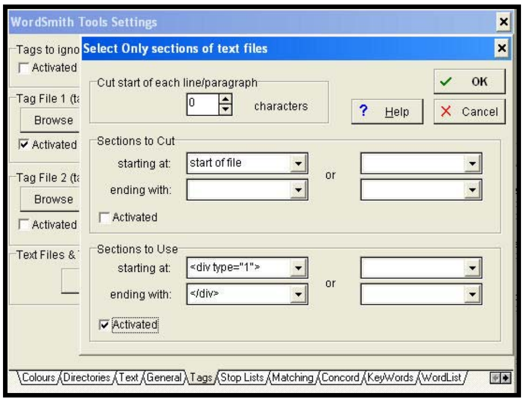
Tag File 2 (tags to be translated):
☐ Activated

Text Files & Tags

☐ Save

Browse
Only if Containing
Only Part of File
OK
Cancel
Help
Notepad
Colours Directories Text General Tags Stop Lists Matching Concord KeyWords WordList

Figure 5. Setting tag list

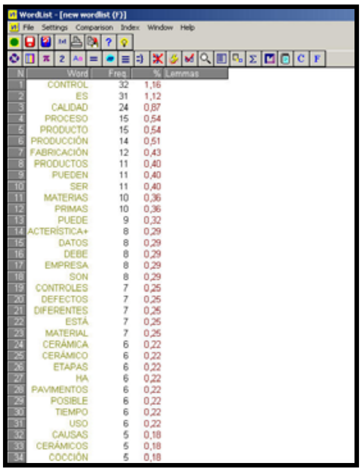

Select Only sections of text files

Cut start of each line/paragraph: 0 characters

Sections to Cut
starting at: start of file or
ending with: or
☐ Activated

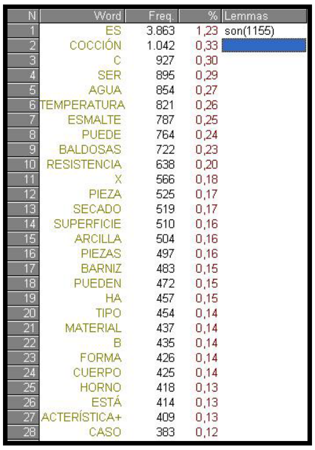
Sections to Use
starting at: <div type="1"> or
ending with: </div> or
☑ Activated

Figure 6. Selecting introduction section


WordList - [new wordlist (F)]

N	Word	Freq.	%	Lemmas
1	CONTROL	32	1,16	
2	ES	31	1,12	
3	CALIDAD	24	0,87	
4	PROCESO	15	0,54	
5	PRODUCTO	15	0,54	
6	PRODUCCIÓN	14	0,51	
7	FABRICACIÓN	12	0,43	
8	PRODUCTOS	11	0,40	
9	PUEDEN	11	0,40	
10	SER	11	0,40	
11	MATERIAS	10	0,36	
12	PRIMAS	10	0,36	
13	PUEDE	9	0,32	
14	ACTERÍSTICA+	8	0,29	
15	DATOS	8	0,29	
16	DEBE	8	0,29	
17	EMPRESA	8	0,29	
18	SON	8	0,29	
19	CONTROLES	7	0,25	
20	DEFECTOS	7	0,25	
21	DIFERENTES	7	0,25	
22	ESTÁ	7	0,25	
23	MATERIAL	7	0,25	
24	CERÁMICA	6	0,22	
25	CERÁMICO	6	0,22	
26	ETAPAS	6	0,22	
27	HA	6	0,22	
28	PAVIMENTOS	6	0,22	
29	POSIBLE	6	0,22	
30	TIEMPO	6	0,22	
31	USO	6	0,22	
32	CAUSAS	5	0,18	
33	CERÁMICOS	5	0,18	
34	COCCIÓN	5	0,18	

Figure 7. Results of introduction section



N	Word	Freq.	%	Lemmas
1	ES	3.863	1,23	son(1155)
2	COCCIÓN	1.042	0,33	
3	C	927	0,30	
4	SER	895	0,29	
5	AGUA	854	0,27	
6	TEMPERATURA	821	0,26	
7	ESMALTE	787	0,25	
8	PUEDE	764	0,24	
9	BALDOSAS	722	0,23	
10	RESISTENCIA	638	0,20	
11	X	566	0,18	
12	PIEZA	525	0,17	
13	SECADO	519	0,17	
14	SUPERFICIE	510	0,16	
15	ARCILLA	504	0,16	
16	PIEZAS	497	0,16	
17	BARNIZ	483	0,15	
18	PUEDEN	472	0,15	
19	HA	457	0,15	
20	TIPO	454	0,14	
21	MATERIAL	437	0,14	
22	B	435	0,14	
23	FORMA	426	0,14	
24	CUERPO	425	0,14	
25	HORNO	418	0,13	
26	ESTÁ	414	0,13	
27	ACTERÍSTICA+	409	0,13	
28	CASO	383	0,12	

Figure 8. Results of body section

6. Conclusion

Our experiment has proved to work with the use of the templates presented here, which shows that their design is correct. They work properly with WordSmith text analyser, and thus, term extraction is improved.

It proved to give flexibility for working with specific sections, because the templates allow to jump from one section to another using simple tools such as a search tool. It can lead to interesting researches about the common position of terms in structured text.

Non-interesting sections can be avoided in order to focus on the ones that tend to have a higher density of terminology. It provides management and control of contextual data. Terminology studies based on the community of speakers that use a specialized language can benefit from the data described in the header.

WordSmith has proved to be a useful tool, althought future work should consider the use of a tool that reads the data type document (DTD) specifications of TEI files, in order to benefit from the semantic information included in the DTD automatically.

7. References

Ahmad, K. & Rogers M. (1997). Corpus Linguistics and Terminology Extraction. In S. E. Wright, G. Budin (eds.), *Handbook of Terminology Management.* Amsterdam & Philadelphia: John Benjamins, vol 2, pp. 725-760.

Alcina, A. (2001). Automatización de Tareas en la Elaboración de Diccionarios Terminológicos. In *Proceedings of Terminologia i documentació. I Jornada de Terminologia i Documentació*. Barcelona: Universitat Pompeu Fabra, pp. 51-60.

Alcina, A. (2005). La Implementación del Concepto de Género Textual en los Corpus Electrónicos para Traductores. In García Izquierdo, I (ed), *El género textual y la traducción*. Berna: Peter Lang, pp. 93-114.

Alcina, A.; Soler, V. & Estellés, A. (2005). Internet como Instrumento para la Documentación en Terminología y Traducción. In Sales Salvador, D. (ed), *Documentarse para Traducir*. Granada: Comares.

Alonso Moro, J. (1989). *Verbos Españoles*. Madrid: Difusón S.L.

Bhatia, V. K. (1993). *Analysing Genre: Language Use in Professional Settings.* London: Longman.

Biber, D.; Conrad S. & Reppen R. (eds.). *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bowker, L. (1996). Towards a Corpus-based Approach to Terminography. *Terminology*, 3(1) pp. 27-32.

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

Faber, P. & Jiménez, C. (eds.) (2002). *Investigar en Terminología*. Granada: Comares.

Garside, R.; Leech G. and McEnery A. (Eds.) (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), pp. 199-220.

Guarino, N. (1995). Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human and Computer Studies,* special issue, 43(5-6), pp. 625-640.

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252-262.

McEnery, T. & A. Wilson. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Miller, G. A. & Fellbaum C. (1991). Semantic Networks of English. In *Cognition*, special issue, 197-229. Reprinted in Levin B. and Pinker, S. (eds.) *Lexical and Conceptual Semantics*. Cambridge, MA: Blackwell, pp. 197-229.

Pérez Hernández, C. (2002) Terminografía Basada en Corpus. In Faber, P. & Jiménez, C. (eds.) *Investigar en Terminología*. Granada: Comares.

Rahtz, S.; Walsh, N. & Burnard, L. (2004). A Unified Model for Text Markup: TEI, Docbook, and beyond. In *Proceedings of XML Europe 2004*. Amsterdam : DeepIX (digital edition).

Reimerink, A. (2002). El Análisis de Corpus para un Fin Práctico: Tendencias en el Uso de los Verbos en la Redacción de Artículos de Investigación. In Faber, P. & Jiménez, C. (eds.) *Investigar en Terminología*. Granada: Comares.

Reppen, R.; Fitzmaurice S. M. & Biber D. (2002). *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Sánchez-Gijón, P. (2004). *L'us de corpus en la traducció especialitzada.* Barcelona: IULA, Universitat Pompeu Fabra.

Sinclair, J.M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Swales, J. M. (1990). *Genre Analysis. English in Academic and Research Settings.* Cambridge: Cambridge University Press.

Tognini-Bonelli, E. (1996). *Corpus Theory and Practice*. Birmingham: TWC.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam & Philadelphia: John Benjamins.

Walsh, N. & Muellner, L. (1999). *DocBook: The Definitive Guide*. O'Reilly & Associates, Inc.