

Parliamentary documents from Spain

Carlos Martin-Dancausa
ETSIT, University of Granada
Periodista Daniel Saucedo Aranda s/n 18071 Granada, Spain
cmdanca@decsai.ugr.es

Maarten Marx
ISLA, University of Amsterdam
Kruislaan 403 1098 SJ Amsterdam, The Netherlands
maartenmarx@uva.nl

ABSTRACT

We created a corpus consisting of all parliamentary documents from Spain since its first legislative period in 1977. The documents were collected from the web page of the Spanish Congress <http://www.congreso.es> and converted into a uniform XML format with extensive metadata in the Dublin Core standard. The collection contains over 50.000 documents with almost 1 million pages having over half a billion tokens. We also collected a complete list of names and biographical data of all members of parliaments during this period. All this data is available for download and will be updated daily. This abstract describes the parliamentary data, the data collection and transformation process and presents some use cases for this corpus. The corpus can be used for corpus-linguistic and political science research, and is suitable for performing scalability tests for XML information systems.

Categories and Subject Descriptors

H.4.m [Information Systems]: Miscellaneous

Keywords

Spanish, Text corpus, Politics, XML

1. COVERAGE AND SIZE OF THE CORPUS

Our aim was to create a corpus containing all digitally available parliamentary documents from the Spanish Congress of Deputies for all the legislative periods¹. A distinction is made between digitally produced and scanned documents.

¹every legislative period is composed of four years of political activity, except for the "first" legislative period (constituent period) that lasts two years

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

LREC2010 Malta

Copyright 2010 ACM ...\$10.00.

The present version of the corpus contains two kinds of documents: Session diaries (Verbatim proceedings) and Official bulletins (Parliamentary documents). Table 1 shows the periods for which digital and scanned data is available on the web for each kind of document.

Table 2 displays information about the size of the corpus. Figure 1 groups the counts per legislative period. We list the following information: the size of the files in text format in Megabytes; the number of documents; the number of pages in the original documents; the number of tokens. We group these numbers for the following two kinds of documents:

The first group are the *verbatim notes* or session diaries of the Congress of Deputies. These can be plenary sessions, sessions of (smaller) committees, united sessions. Even though the texts are edited and transcribed to be read, they are accounts of spoken language. The second group, called *parliamentary documents*, are composed of the descriptions of law proposals, law projects, treaties, international agreements, etc.

2. AVAILABILITY AND IPR

The corpus is available for download at <http://www.politicalmashup.nl/SpanishParliament>. The Congress of Deputies allows to work with all these public PDF documents from the webpage but it is forbidden to use them with commercial purposes so we are not aware of copyright restrictions on the material. If you use the corpus, please send an email to maartenmarx@uva.nl.

3. TECHNICAL DESCRIPTION

3.1 Description of the data format

Every document in the corpus is an UTF-8 encoded XML file which is valid with respect to the Relax NG schema developed for the parliamentary information in Dutch [3]. Here we briefly describe the structure of the documents. The root of each document has three children:

meta this element contains meta-information of the document described using the 15 elements from the Dublin Core Metadata Element Set Version 1.1²;

header this element contains textual data extracted from the source-text which may be used for displaying purposes;

²<http://dublincore.org/documents/dces/>

Kind of document	Scanned	Digital
Plenary Sessions	From 1977-07-13 (Constituent Leg. Per. - V Leg. Per.)	From 1996-3-27 (VI Leg. Per. - IX Leg. Per.)
Official Bulletins	From 1977-07-12 (Constituent Leg. Per. - V Leg. Per.)	From 1996-4-3 (VI Leg. Per. - IX Leg. Per.)

Table 1: Availability of parliamentary data in Spain.

	Session Diaries			Official Bulletins		
	# Documents	Pages	Words	# Documents	Pages	Words
Const. Leg. Per.	196	8182	5411762	224	4864	2475881
Leg. I	410	21152	14405148	10366	36123	12998073
Leg. II	799	27995	24390594	2661	29549	16520591
Leg. III	707	28212	26008882	2994	44616	22516414
Leg. IV	931	33528	30515127	3694	52855	25037065
Leg. V	847	30227	27363942	3621	46021	21665665
Leg. VI	1121	40016	37263530	5674	71212	38505024
Leg. VII	1226	46570	40855054	5169	125414	59125517
Leg. VIII	1313	46112	41504131	6900	159104	78343241
Leg. IX	470	13953	12633288	1544	62237	34535336
Total	8020	295947	260351458	42847	631995	311722807

Figure 1: Statistics of the Spanish corpus from the different legislative periods

Subcorpus	Mb text	# Documents	# Pages	# Tokens
Verbatim proceedings	17387	8020	295947	260351458
Parliamentary documents	20516	42847	631995	311722807
Total	37903	50867	927942	572074265

Table 2: Number of documents, pages and tokens for parliamentary documents, verbatim proceedings and the complete corpus.

text this element contains the complete text of the source document. Each **text** element has one or more **page** elements (corresponding to physical pages of the document), which in turn are divided in one or more **p** (for paragraph) elements.

Within the **text** element there is a strict separation between content and metadata. All metadata is stored in attributes. All text is contained in the **p** elements. The attributes of the **page** and **p** elements give unique names, and contain provenance information. Both have an obligatory **docno** attribute whose value is unique in the corpus and together with a suitable namespace a URN [5]. This conforms to the recommendations of publishing eGovernment material as set out by the eGov working group of the W3C [1].

3.2 Description of the data collection and processing

The web page of the Congress of Spain <http://www.congreso.es> contains the information we want to collect in one corpus. We have developed different scripts in Perl, which are automatic processes in charge of downloading all the metadata and the PDF documents. We describe here the main steps.

Examining the search engine on the parliaments web page, we have observed that the URL addresses of the result pages contain all the parameters introduced in the query so if we use this URL address with all the different parameters in which we are interested in, we can download all the lists of documents for every legislative period in an automatic way. Then, it is turn to begin to work with these lists in HTML format getting the URL link to download the PDF documents and some information interesting for our corpus. This information consists of a list of attributes:

- Name of the file.
- Date of the publication.
- Kind of document.
- Number of document.
- Number of pages.

The number of pages does not appear in the web page but we have used **pdfinfo**³ which retrieves information about PDF documents like the number of pages.

Having collected the PDF documents and the metadata, we transform them into the XML format using the transformation described in [3].

3.3 Deputies details

Apart from the documents of the Congress of Deputies, we want to get personal details of the deputies. This information is useful for named entity reconciliation [2] in the verbatim proceedings. Biographies of all deputies are available from the web page of the Spanish parliament. We downloaded the biographies using a similar method as described before.

We collected the following information for each deputy: Name; Surnames; Political party; Legislative Period which it belongs; Start date in the Congress; End date in the Congress (if it has retired); Link to the photograph; Link

³This is part of the Xpdf software, see <http://www.foolabs.com/xpdf>.

to the Congress web page. This data is available in in csv format.

An initial experiment with using this database for recognizing and disambiguating speakers in verbatim proceedings showed that the extra information is useful and needed, in particular for the scanned documents which contain OCR-errors.

4. A LOOK AT THE DATA

As we can see in Figure 2, there is a large difference between the number of official bulletins of the first (I) Legislative Period and the number of official bulletins of the other legislative periods. This is due to a change in the filing system. In later legislative periods several items which were published before as separate documents are grouped.

Figure 3 shows word counts of a number of words per legislative period. We have looked at the following words:

ETA This is a Spanish terrorist organization, which has committed a lot of terrorist attacks. The worst terrorist attack in Spain was in Madrid in the VIII Leg. Per. and ETA was an important suspect. This is reflected in the sharp rise in the number of appearances of ETA during this legislative period.

Crisis Nowadays, it is a very common term due to the Spanish situation: Recession, Unemployment, Minimum salaries, etc. It is also possible to distinguish different crisis periods in the graphic.

Franco

Irak

Palomares This is the place where two planes crashed in 1966 and an atomic bomb fell into the sea without exploding. This problem was an important matter during several years because people were afraid of taking a bath in the sea becoming a frequent topic in the Congress in the I Legislative Period.

Studying the graphics, we have concluded there are fewer appearances of the different words in the first legislative periods. Most probably this is because all the documents from the first legislative periods are scanned and we find several mistakes in the transcriptions of the words from the scanned documents to the text documents.

5. CONCLUSIONS AND FUTURE WORK

The objective of our work is to develop a corpus of all official documents of the Parliament in Spain since its beginnings in 1977. The session diaries and the official bulletins from Congress of Deputies included in the present version of the corpus are a good starting point.

This corpus may be used in a XML retrieval system [4] similar to <http://theyworkforyou.com>.

The Congress of Deputies has a multimedia collection with all the videos of session diaries segmented at the speaker level. When the verbatim proceedings are also segmented at the speaker level these can be linked and a powerful video search engine results.

For linguistic research into the Spanish language it would be good to extend the corpus with parliamentary data from other Spanish speaking countries. Argentina for instance has clearly structured information at <http://www.diputados.gov.ar/>.

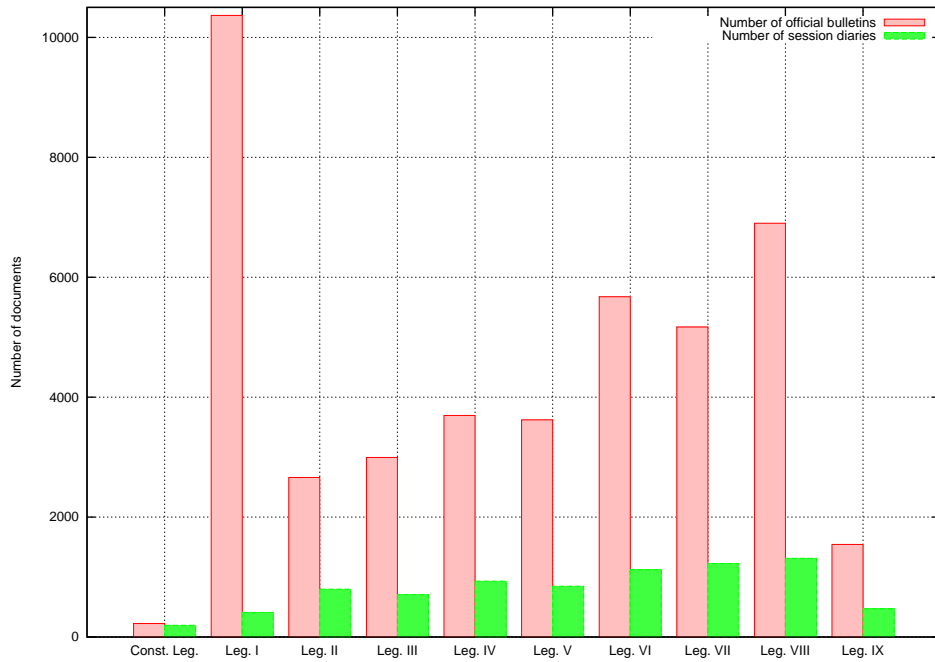


Figure 2: Number of documents of the Spanish corpus from the different legislative periods

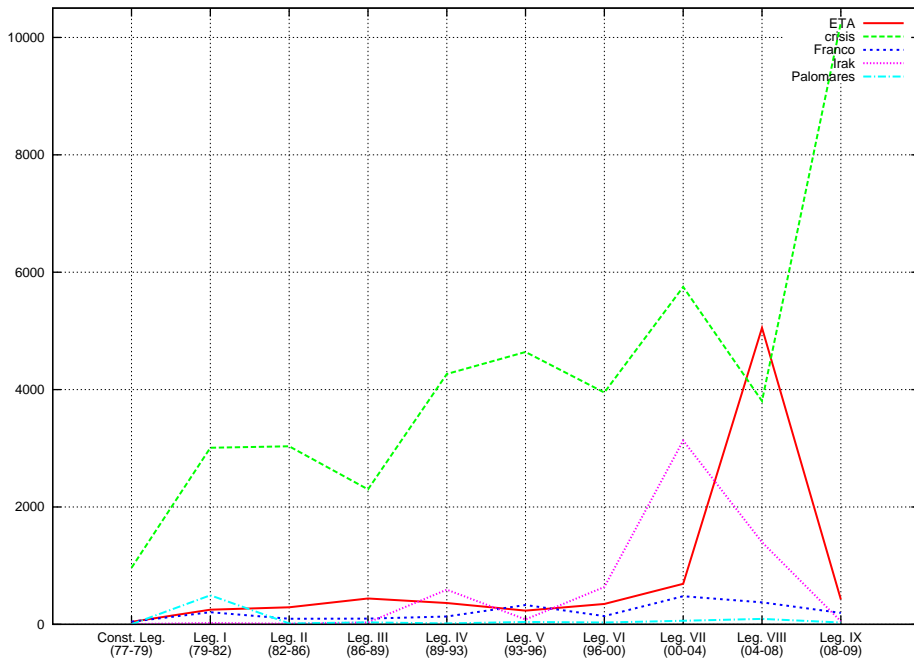


Figure 3: Number of appearances of several special words

Acknowledgements

Maarten Marx acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under the FET-Open grant agreement FOX, number FP7-ICT-233599.

6. REFERENCES

- [1] D. Bennet and A. Harvey. Publishing open government data (W3C Working Draft 8 September 2009). <http://www.w3.org/TR/gov-data/>, 2009.
- [2] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proc. SIGMOD*, pages 85–96, 2005.
- [3] M. Marx and A. Schuth. DutchParl A Corpus of Parliamentary Documents in Dutch. <http://politicalmashup.nl/DutchParl>.
- [4] B. Sigurbjörnsson. *Focused information access using XML element retrieval*. PhD thesis, University of Amsterdam, 2006.
- [5] W3C/IETF URI Planning Interest Group. URIs, URLs, and URNs: Clarifications and Recommendations 1.0. W3C Note 21 September 2001, 2001. <http://www.w3.org/TR/uri-clarification>.