DutchParl
A Corpus of Parliamentary Documents in Dutch

Maarten Marx and Anne Schuth
ISLA, University of Amsterdam
Kruislaan 403 1098 SJ Amsterdam, The Netherlands
maartenmarx@uva.nl aschuth@science.uva.nl

ABSTRACT

A corpus called DutchParl is created which aims to contain all digitally available parliamentary documents written in the Dutch language. The first version of DutchParl contains documents from the parliaments of The Netherlands, Flanders and Belgium. The corpus is divided along three dimensions: per parliament, scanned or digital documents, written recordings of spoken text and others. The digital collection contains more than 800 million tokens, the scanned collection more than 1 billion.

All documents are available as UTF-8 encoded XML files with extensive metadata in Dublin Core standard. The text itself is divided into pages which are divided into paragraphs. Every document, page and paragraph has a unique URN which resolves to a web page. Every page element in the XML files is connected to a facsimile image of that page in PDF or JPEG format. We created a viewer in which both versions can be inspected simultaneously. A search-engine for the complete collection is available online.

The corpus is available for download in several formats. The corpus can be used for corpus-linguistic and political science research, and is suitable for performing scalability tests for XML information systems.

Keywords

Dutch, Text corpus, Politics, XML

1. INTRODUCTION

The main reason to create the corpus is to provide one portal from which these documents are accessible both in their original official version (in PDF format), and in a uniform XML format with extensive metadata [2]. The corpus was designed to be useful as a data set in all possible scientific disciplines. E.g., it can be used for (comparative) corpus-linguistic and political science research and as a test-set for information-theoretic experiments. This distinguishes DutchParl from EuroParl [1] which is developed for research in Statistical Machine Translation. The corpus was developed following the guidelines set out in [2].

DIR2010 Nijmegen

How to get the corpus?

The corpus is available for download at

`http://politicalmashup.nl/DutchParl`

A full version of this paper is also available there. We are not aware of copyright restrictions on the material. If you use the corpus, please sent an email to `maartenmarx@uva.nl`.

2. COVERAGE AND SIZE OF DUTCHPARL

Spatial and temporal coverage.

Parliamentary documents in the Dutch language are produced in the following locations: Belgium (Flemish parliament, and the Belgian federal parliament), European Union, Suriname and The Netherlands. The present version of DutchParl does not yet contain data from the EU nor from Suriname.

The periods for which data is available differ per source. Table 1 lists the periods for which digital and scanned data is available on the web for each source (measured in September 2009). This is exactly the data available in DutchParl.

Subcorpora.

The corpus can be divided into many subcorpora. This is facilitated by the uniform metadata using a controlled vocabulary. In the description below we partition the data along three dimensions. First by source: Belgium, Flanders and The Netherlands. Secondly, digitally produced documents are separated from scanned and OCR-ed documents. The latter contain noise in the form of wrongly recognized characters, mistakes in paragraph splitting, non UTF-8 characters, or simply no extractable text.

A special subset of the parliamentary documents are the verbatim notes of sessions of parliament. Even though the texts are edited and transcribed to be read, they are accounts of spoken language. For this reason, we present details both for the complete collections and for the verbatim notes separately.

Size of DutchParl.

Table 2 displays information about the size of the subcorpora. We note that the documents from the Belgian parliament are bilingual, with text in Dutch and French interspersed in many different ways.

Source	Digital	OCR-ed	Planned
Belgium	From 1999-07-01	-	1844–1999 is scanned
Flanders	From 1995-10-17	1971-12-07 to 1995-10-17	-
The Netherlands	From 1995-01-01	1917-01-01 to 1995-01-01	1814–1917 available in 2010

Table 1: Availability of parliamentary data in the Dutch language.

Subcorpus	Mbyte text	# Documents	# Pages	# Tokens
Belgian Federal	800	3.901	216.522	129.085.483
Flanders	454	5.470	161.881	72.958.408
Netherlands	4.331	198.433	1.594.845	684.932.669
Flanders OCR	146	1.018	34.867	23.924.567
Netherlands OCR	7.043	328.722	1.701.130	1.003.555.596

Subcorpus	Mbyte text	# Documents	# Pages	# Tokens
Belgian	502	3.462	137.366	81.086.575
Flanders	311	3.799	93.591	50.715.218
Netherlands	781	21.604	137.610	131.681.453
Flanders OCR	142	932	33.147	23.378.215
Netherlands OCR	2.644	12.796	383.863	402.657.396

Table 2: Number of documents, pages and tokens for the complete corpus (top) and only for verbatim notes of parliamentary and committee sessions (bottom).

	NL-DIGITAL	NL-SCAN	Flanders DIGITAL	Flanders SCAN	BE-federal
Total number of words	102870201	329540359	38629223	17120704	41152224
Unique words	353677	1963712	258304	184945	245447
Words occurring just once	149719	1311243	118992	91889	102093
Words occurring more than once	203958	652469	139312	93056	143354
Words occurring at least 4 times	130008	370932	88518	57277	90911
Words occurring at least 20 times	55054	134735	36413	22945	37250

Table 3: Token counts; all data (top) and verbatim notes of parliamentary sessions (bottom).

Number of tokens.

Table 3 presents figures on the number of tokens occurring in the different subcorpora. Again we make a distinction between digital and scanned documents and present the numbers for the spoken texts separately.

3. CONCLUSIONS AND FUTURE WORK

This work started out as a challenging data integration project: Can we collect and bring together under one uniform schema parliamentary data from different countries, produced in different periods of time, and available in different formats? DutchParl showed that we partly succeeded. We created a rich metadata schema based on Dublin Core standards. However, it is not always easy or possible to collect meaningful data for all fields (we did not manage for Belgium Federal). Also, even after many tries and promises, we did not receive the data from Suriname. A hard problem is checking completeness. Even if we are confident that we downloaded all material available on the web, we cannot be sure that we have all material. It is difficult to find reliable independent listings of material.

We paid extra care to providing provenance information. Because we assigned corpus unique ID's to every paragraph, page and document, specific referencing of material (common in the social sciences) is possible using hyperlinks. The connection of the data in XML with the original official publications is quite specific and convenient because we provide a facsimile image of every page.

Future challenges include 1) keeping the corpus daily up to date, 2) managing the data in an XML database managment system, 3) scaling to other countries, 4) linking the data with other datasets, e.g. bibliographies of MP's, 5) performing text analytics on noisy data, 6) machine translation, 7) create a search system.

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.

4. REFERENCES

[1] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proc. MT summit*, volume 5, 2005.

[2] M. Wynne. Archiving, distribution and preservation. In M. Wynne, editor, *Developing Linguistic Corpora: a Guide to Good Practice*, pages 71–78. Oxford: Oxbow Books, 2005. Available online from http://ahds.ac.uk/linguistic-corpora [Accessed 2009-07-01].