EEOP2012: Exploring and Exploiting Official Publications

Workshop Programme

09:00-09:05	Welcome and introduction by Steven Krauwer
09:05-09:50	**Invited talk:** Maarten Marx *Open Official Documents: Requirements and Opportunities*
09:50-10:10	Michael Rosner and Andrew Attard *Intelligent Exploitation of Local Government Resources*
10:10-10:30	Maria Palmerini, Ruben Cerolini, Giulio Santini and Francesco Cutugno *From Recording to Retrieving: A Proposal of a Complete System for Semi-automatic Reporting for Local and National Governments*
10:30-11:00	Coffee break
11:00-11:20	Vidas Daudaravicius *Automatic Multilingual Annotation of EU Legislation with Eurovoc Descriptors*
11:20-11:40	Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru *GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain*
11:40-12:00	Oliver Mason, Aleksander Trklja and Dominik Vajn *Requirement Extraction from Transport Policy Documents*
12:00-12:50	Open discussion *What are possible actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level?*
12:50-13:00	Winding up and closing by Steven Krauwer

Editor

Steven Krauwer	Utrecht University / CLARIN ERIC

Workshop Organizers and Programme Committee

Steven Krauwer	Utrecht University / CLARIN ERIC
Ralf Steinberger	European Commission – Joint Research Centre (JRC)
Arjan van Hessen	University of Twente / CLARIN-NL
Nicoletta Calzolari	CNR-ILC / ELRA
Hans Uszkoreit	DFKI / META-NET

Table of contents

Author Index

Introduction

Steven Krauwer, Nicoletta Calzolari, Arjan van Hessen, Steven Krauwer, Ralf Steinberger, Hans Uszkoreit

The EEOP2012 workshop is dedicated to the exploitation and exploration of official publications in digital format, both at the international level (often multilingual) and at the national level (mostly monolingual, but in some cases multilingual as well). These publications can be in written, spoken or visual form or combinations thereof (e.g. written proceedings of parliaments, legislative documents, audio or video recordings of parliament sessions, simultaneous translations by interpreters or in sign language).

The workshop covers the whole lifecycle of these publications, ranging from acquisition, annotation, instrumentation, exploration of data and content, exploitation of data and content to support research and the development of tools and applications.
The main objectives of the workshop are:

- To create awareness of the importance of official publications by showing the research and development possibilities they offer;
- To share results, experiences and problems emerging from work on a variety of corpora, modalities and languages;
- To identify actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level.

Official publications can be of tremendous importance for the research communities interested in human language technology (in the broadest possible sense) and for the communities interested in linguistics, psychology, history, social sciences and political sciences because they have a number of specific characteristics that make them different from other language resources:

- If they exist in digital form they are normally public and free;
- They grow continuously;
- They are often multilingual and parallel;
- They lend themselves for exploitation (as training material for tools and sometimes possibly even for niche applications);
- They lend themselves for exploration to support linguistic studies, studies about human behaviour, about changes in society, attitudes, and many other possible research topics in the humanities and social sciences;
- Because of their comparability they lend themselves for porting technologies, methods and expertise between languages;
- They lend themselves for educational purposes for technologists, linguists and other scholars.

Primary audience of this workshop is:

- Language and speech technology researchers from academia and industry;
- Humanities and social sciences scholars with an interest in digital methods;
- Educators in these fields.

Additional beneficiaries, not necessarily present at LREC 2012:

- Professionals interested in analysing political behaviour or processes (e.g. journalists, policy makers, policy watchers);
- Parties interested in providing or exploiting such analysis tools on a commercial basis;
- Translation studies scholars;
- Comparative linguists.

Workshop home page: http://www-sk.let.uu.nl/EEOP2012
Contact: s.krauwer@uu.nl

Acknowledgement: This workshop was supported by the European Commission – Joint Research Centre (JRC)

Exploring and exploiting official publications: requirements analysis*

Maarten Marx

ISLA, University of Amsterdam
Science Park 904
1098 XH Amsterdam, The Netherlands
maartenmarx@uva.nl

Abstract

We address the problem of publishing parliamentary proceedings in a digital sustainable manner. We give an extensive requirements analysis, and based on that propose a uniform XML format. We evaluated our approach by collecting and automatically processing proceedings from six parliaments spanning almost 200 years in total. Most of this data is real legacy data consisting of scanned and OCRed documents. The approach scales well and produces high quality data.

Keywords: Knowledge Representation, XML, Parliamentary Proceedings, Linked Data

1. Introduction

Many democratic countries recently witness a rise in publishing governmental data on the web. Data is made available by different institutions: local and central governments, commercial publishing houses and non-commercial initiatives like `http://theyworkforyou.com`. Because many countries create very similar data, e.g. constitutions, tax laws and parliamentary proceedings, it is beneficial to try to standardize the format in which data is published. A standardized format has many advantages: comparative studies are facilitated, software can be exchanged and universally applied, emergence of best practises and a community of expert users. Several initiatives to standardization are or have been taken. An active group is the W3C working group on Egovernment which published two recommendations (Alonso et al, 2009; Bennet and Harvey, 2009).

This paper focuses on one particular dataset which is produced in almost every democratic country and which has great appeal to both the general public, the media, and the scientific community: the proceedings of the plenary meetings of parliament. These are very well structured, verbatim notes of everything that is being said and that happens during plenary sessions.

An important aspect of parliamentary proceedings is their longitudinal character. In many states, the proceedings are collected for more than 100 years. Numerous states are currently digitizing their legacy using scan and OCR techniques. This trend gave rise to our research question:

> What is the best data format for publishing both legacy and current parliamentary proceedings in a digital sustainable manner?

Our main results are recommendations for representation schemas for the most important data format for publishing open government data, XML (Alonso et al, 2009; Bennet and Harvey, 2009; Berners-Lee, 2006). We evaluated the effectiveness of our representation by collecting almost 200 years of proceedings from five parliaments and transforming these into the common representation.

The paper is organized as follows. Section 2. contains an extensive analysis of the requirements on a good representation. Section 3. gives a detailed description of the XML format we have developed. We end with conclusions. For an evaluation of the proposed approach we refer to (Marx and Schuth, 2010).

Methodology We used the following methodology for arriving at the requirements on the representation of parliamentary documents. We surveyed existing comparative scientific research based on parliamentary proceedings and distilled desiderata. We investigated current representations and information systems in six states[1], and we took the recommendations for publishing governmental data as linked data from the W3C (Alonso et al, 2009; Bennet and Harvey, 2009). This resulted in a large wish list which no country in our survey could yet satisfy.

We then investigated which parts of the wish list could be fulfilled effectively with fully automated processes. The main criterion used was *scalability*: techniques tailor made for specific time periods in specific states were mostly discarded. We used techniques from information extraction and retrieval (Manning et al., 2008; Liu, 2007; Rahm and Do, 2000) to *automatically* convert currently available data into the desired formats. Automatic conversion is an essential requirement because of the vast amounts of legacy data around. Usually this legacy data is only available as scanned and OCRed copies of printed versions.

The dataset reminds us of the youth of the digital age. The eldest proceedings in our survey which are available in an original digital format are from 1995.

We implemented the chosen techniques and tested them by converting proceedings into machine processable format for six parliaments covering almost 200 years. We then evaluated the accuracy of these techniques.

2. Requirements assessment

To answer our research question we collect requirements on publishing parliamentary data from four different sources. We first investigate the intrinsic qualities of the data itself.

* This paper is a shortened version of (Marx et al., 2010).

[1] Austria, Belgium, Germany, The Netherlands, Spain and the UK.

Then we survey typical scientific research done on parliamentary data and extract requirements from that. Thirdly, we look at W3C recommendations on publishing governmental data. We finish with a list of features collected from our survey websites publishing parliamentary data.

2.1. Intrinsic qualities

The most valuable characteristic of a collection of parliamentary proceedings is its longitudinal nature. The collection consists of periodic measurement points conducted in a uniform and consistent manner over a (possibly very long) period of time. The data is thus well suited for temporal comparisons. Also, measurements are rather similar across states which facilitates cross-national comparative studies, common in the political sciences.
The collection is a record of spoken language with very rich metadata. For every word spoken in parliament, the following facts are known, and can be extracted from the written proceedings:

1. when it was said,
2. who said it,
3. in what function,
4. speaking on behalf of which party,
5. in which context, and
6. who was actively present during the speech act.

These features enable all kind of groupings and comparisons. Findings in different states may also be compared. It is desirable that a representation makes these six features machine processable.

2.2. Scientific research

We distinguish qualitative and quantitative research, as each comes with their own requirements.
Because of their longitudinal nature, parliamentary proceedings are important data for historical research. It is a goldmine for historic-linguistic and etymological research looking for first (spoken) occurrences of terms. This qualitative research requires powerful search capabilities (e.g. using wildcards for characters to allow for OCR-errors), fast access to processed and raw data (in this case usually the OCRed text and the scanned images, conveniently linked), and the ability to make precise references into the source material (comparable to the very fine-grained reference system of the Bible).
Fields as political science, sociology, communication science and content analysis additionally use quantitative methods to study large amounts of textual data (Lazer et al., 2009). Modern text analytics techniques from the fields of information retrieval (Manning et al., 2008) and web data mining (Liu, 2007; Hillard et al., 2007) are applied here. Examples include agenda-setting research (McCombs and Shaw, 1972), research correlating parts of the political spectrum with specific (e.g. populist) language (Jagers, 2006), and trend detection in media and parliament (Roggeband and Vliegenthart, 2007). This research uses exactly the six features from the previous section.

The Text Encoding Initiative (TEI `http://www.tei-c.org`) publishes XML schemas for various kind of publications, but not for parliamentary proceedings.

2.3. W3C recommendations

The W3C created three notes on publishing government data (Alonso et al, 2009; Bennet and Harvey, 2009; Berners-Lee, 2009). The main points are:

- make data both machine and human readable;
- link data, make data linkable, provide permanent identifiers for each government object and data item;
- provide metadata using common standards (e.g. Dublin Core);
- make the data as easy to reuse (e.g. in mashups) as possible.

Tim Berners Lee (Berners-Lee, 2009; Berners-Lee, 2006) emphasizes the fact that government data should be published as *linked data*. This means that it is open (expressed in non-proprietary formats; XML and RDF are preferred), modular (data can be combined with other pieces of data), and scalable.
According to (Alonso et al, 2009), "much public sector information was and is stil being published using proprietary formats or in ways that create barriers of use for various interested parties". Potential benefits of open and linkable data include multiple views (e.g., list everything being said by MP *X*), reuse of information, improved web search and data integration.
(Alonso et al, 2009) also mentions provenance and trust explicitly, here in connection with mashups. The data-format should make it very easy to refer and return to the original data source, both for machines and humans. Making data linkable using permanent identifiers is also recognized by the OECD who use Digital Object Identifiers (DOI's) for permanent links (Green, 2009).
The eight principles published by *The Open Government Group* (`http://www.opengovdata.org`) and reproduced in Table 1, neatly summarize the W3C recommendations.

2.4. Best practices

We analysed the websites of six parliaments and two independent foundations providing access to parliamentary information. They are listed in Table 2. Here we provide a list of best practices that we found and that are relatively easy to implement. Nonetheless, none of these points was present at the majority of the sites.

- Publish extensive metadata to each file, appropriately linked and in a common format.
- Make the status of the Intellectual Property Rights of the data clear and easy to find.
- Publish fast, thus also publish non-definitive versions (UK).

1. Complete All public data is made available. Public data is data that is not subject to valid privacy, security or privilege limitations.

2. Primary Data is as collected at the source, with the highest possible level of granularity, not in aggregate or modified forms.

3. Timely Data is made available as quickly as necessary to preserve the value of the data.

4. Accessible Data is available to the widest range of users for the widest range of purposes.

5. Machine processable Data is reasonably structured to allow automated processing.

6. Non-discriminatory Data is available to anyone, with no requirement of registration.

7. Non-proprietary Data is available in a format over which no entity has exclusive control.

8. License-free Data is not subject to any copyright, patent, trademark or trade secret regulation. Reasonable privacy, security and privilege restrictions may be allowed.

Compliance must be reviewable.

Table 1: Open Government data principles, C. Malamud et. al. 8 December 2007, `http://resource.org/8_principles.html`.

Austria	`http://www.parlinkom.gv.at`
Belgium	`http://www.dekamer.be`
European Union	`http://www.europarl.europa.eu/activities/plenary/home.do`
Flanders	`http://www.vlaamsparlement.be`
Germany	`http://www.bundestag.de`
The Netherlands	`http://parlando.sdu.nl/cgi/login/anonymous`
Spain	`http://www.congreso.es`
MySociety.org (UK)	`http://theyworkforyou.com`
PoliticalMashup (NL)	`http://polidocs.nl`

Table 2: Parliamentary information websites that were surveyed.

- Link named entities occurring in the proceedings. In Austria, members of parliament and government who are speaking are linked to their biographical page. Also numbers referring to laws or dossiers are linked to their pages.
- Put timestamps in the proceedings in order to allow alignment with audiovisual material. Austria does this at the start of every new speaker.
- Publish in well formed XML (XHTML) in order to allow for machine processing.
- Publish in coherent wholes. E.g. publish the proceedings of one day together in one file.
- Publish multiple views which link into the original and to other documents. For instance, publish on the biographical page of each member of parliament a list of all her speeches. Each speech should be linked to its place in the debate to provide context (Austria, EU).

All of these best practices are instances of the principles from Table 1.

2.5. In summary

In summary we can distill four main points:

- add metadata in a uniform standard format;
- publish in XML with links to the original sources;
- make each object linkable by unique permanent identifiers;
- make entities explicit in the representation and link them appropriately.

The next section formalizes these points as constraints on the XML representation.

3. Representation in XML

This section contains the XML schema in which we represented parliamentary proceedings. (Gielissen and Marx, 2009) contains the following high level schema for documents describing the proceedings of one day:

```
meeting        ⟶  (topic)+
topic          ⟶  (speech | stage-direction)+
speech         ⟶  (p | stage-direction)+
p              ⟶  (#PCDATA | stage-direction)*
stage-direction ⟶  (#PCDATA).
```

All elements contain metadata stored in attributes. The speech elements contain attributes specifying the name, the party and the function in parliament of the speaker. This purely semantical representation satisfies the requirements of Section 2.1..

Because of its semantical nature, special purpose extraction scripts need to be created for each parliament. Moreover, for each change in layout or organization of the proceedings, adaptations of the extraction scripts are called for. Thus this approach does not scale well.
An alternative layout oriented representation does not have this scalability problem, but still allows further processing which extracts the semantical information. We now describe that in detail.
Every document is a UTF-8 encoded XML file which is valid with respect to the Relax NG schema, available from the authors. We briefly describe the structure of the documents. The root element `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 1.1[2];

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

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 XPath expression `doc('file.xml')//text//text()` will return the complete text of the source document.
The attributes of the `page` and `p` elements contain provenance information (Hartig, 2009). The `root`, `page` and `p` elements have an obligatory `docno` attribute whose value is unique in the corpus. Each `page` also has an obligatory `imageref` attribute which points to a facsimile image of that particular page (these can be in PDF or JPEG format). All other attributes are optional. We briefly list them:

originalpagenr an integer denoting the page number of the page in the original document. This is extracted from the text using a special pattern. If the confidence in the extracted value is too low a '-' is given as a value.

class Its value is either "header" or "footer". Determined from the text using heuristics.

top and left Integers denoting the position of the upper left hand corner of the bounding box of the paragraph. The length of each page is normalized to 1000 units.

fulltextref and wordcoordinatesref These are two URLs referring to files which are specific for the Dutch OCR-ed part of the collection.

[2] `http://dublincore.org/documents/dces/`

Dublin Core metadata Metadata is described in a uniform way for all sub-collections using the 15 Dublin Core properties. A number of elements obtained a fixed value for the complete collection. We briefly discuss the others. `dc:coverage` indicates the country or region of the parliament. `dc:date` refers to the date of the document. This is often hard to determine, and in many cases not available. For documents of `dc:type` "Written Questions" the `dc:date` element is subdivided into the date of the question, the date of the answer and the difference between these two in number of days, whenever these could be obtained from the metadata.
`dc:description` and `dc:title` are free text describing the document.
`dc:publisher` contains the URL of the website from which the data is harvested. `dc:rights` contains the name of the parliament which produced the document. `dc:identifier` contains the URL of the present XML file. `dc:source` contains URLS to the text source and (if available) the source of the metadata.
`dc:type` indicates the kind of parliamentary documents. We distinguish two types: *Verbatim Proceedings* contain the meeting notes of plenary sessions of the parliament; *Written Questions* contain written question of members of parliament to members of the government and the answers. All other documents obtain type *Parliamentary Documents*.
The properties `dc:relation` and `dc:subject` contain semantic information which is usually not available and needs to be extracted from the text. These are not used yet.
We tried to restrict the fields as much as possible. With the data-type restrictions this may lead to validation errors due to typos or mistakes in the data. For instance, the string `2008-04-31` will not be accepted as being of type `xsd:date`, because that date does not exist.

4. Conclusions

We addressed the problem of publishing parliamentary proceedings in a digital sustainable manner. We gave an extensive requirements analysis, and based on that proposed a uniform XML format. An extensive evaluation (Marx and Schuth, 2010) showed that the approach scales very well and produces high quality data.
Although the paper only discussed parliamentary proceedings we believe that both the findings and the used methodology are applicable to other governmental datasets. We have successfully applied our techniques to political speeches and written questions and answers. The thus obtained datasets were used for several applications ranging from political search systems (Nusselder et al., 2009) and electoral advice systems (Jijkoun et al., 2007) to debate summarization systems (Kaptein et al., 2009; Marx, 2009).

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. This research

was supported by the Netherlands organization for Scientific Research (NWO) under project number 380-52-005 (PoliticalMashup).

5. References

J. Alonso et al. 2009. Improving access to government through better use of the web. W3C Interest Group Note 12 May 2009 http://www.w3.org/TR/egov-improving/, May.

D. Bennet and A. Harvey. 2009. Publishing open government data (W3C Working Draft 8 September 2009). http://www.w3.org/TR/gov-data/.

Tim Berners-Lee. 2006. Linked data. http://www.w3.org/DesignIssues/LinkedData.html.

Tim Berners-Lee. 2009. Putting government data online. http://www.w3.org/DesignIssues/GovData, June.

T. Gielissen and M. Marx. 2009. Exemelification of parliamentary debates. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop (DIR 2009), Twente, The Netherlands*, pages 19–25.

T. Green. 2009. We need publishing standards for datasets and data tables. Technical report, OECD Publishing White Paper. http://dx.doi.org/10.1787/603233448430.

O. Hartig. 2009. Provenance Information in the Web of Data. In *Proc. of the Linked Data on the Web Workshop at WWW*.

D. Hillard, S. Purpura, and J. Wilkerson. 2007. Computer-assisted topic classification for mixed-methods social science research. *Journal of Information Technology & Politics*, 4(4):31–46.

J. Jagers. 2006. *De stem van het volk. Populisme als concept getest bij Vlaamse politieke partijen*. Ph.D. thesis, Universiteit Antwerpen.

Valentin Jijkoun, Maarten Marx, Maarten de Rijke, and Frank van Waveren. 2007. Electoral search using the verkiezingskijker: an experience report. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 1155–1156, New York, NY, USA. ACM Press.

R. Kaptein, M. Marx, and J. Kamps. 2009. Who said what to whom? Capturing the structure of debates. In *Proceedings SIGIR '09*, pages 831–832.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. 2009. Computational social science. *Science*, 323(5915):721–723.

B. Liu. 2007. *Web Data Mining*. Springer.

Ch. Manning, P. Raghavan, and H. Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

M. Marx and A. Schuth. 2010. DutchParl. A Corpus of Parliamentary Documents in Dutch. In *Proceedings Language Resources and Evaluation (LREC) 2010*, pages 3670–3677.

M. Marx, N. Aders, and A. Schuth. 2010. Digital sustainable publication of legacy parliamentary proceedings. In *dg.o '10: Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, pages 99–104. Digital Government Society of North America. http://portal.acm.org/citation.cfm?id=1809874.180989

M. Marx. 2009. Long, often quite boring, notes of meetings. In *ESAIR '09: Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pages 46–53. ACM.

M. McCombs and D. Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36:176–187.

A. Nusselder, H. Peetz, A. Schuth, and M. Marx. 2009. Helping people to choose for whom to vote. a web information system for the 2009 European elections. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management, CIKM 2009, Hong Kong, China, November 2-6, 2009*, pages 2095–2096. ACM.

E. Rahm and H.H Do. 2000. Data cleaning: Problems and current approaches. *IEEE Techn. Bulletin on Data Engineering*, 23(4).

C. Roggeband and R. Vliegenthart. 2007. Divergent framing: The public debate on migration in the Dutch parliament and media, 1995-2004. *West European Politics*, 30(3):524–548, May.

Intelligent Exploitation of Local Government Resources

Mike Rosner, Andrew Attard
Faculty of ICT, Dept. Intelligent Computer Systems
Univerisity of Malta, Msida MSD2080 MAlta
E-mail: mike.rosner@um.edu.mt, andrew.attard@um.edu.mt

Abstract

Malta is divided into sixty-eight local councils each contributing to the most basic form of local government. Several meetings take place during which the councillors gather to discuss the maintenance and embellishment of the locality, each of which are noted down in Maltese. This paper concerns a corpus of local government documents. We suggest an approach to the problem of developing an intelligent browsing system that offers improved access to the information, for example to assist local councils in decision making, or to give members of the public more transparent way to browse local council documentation.

Keywords: Govenment corpus; intelligent catalogue system

Acknowledgement: The development of this paper has been partially supported by the Seventh Framework Programme and the ICT Policy Support Programme of the Euro- pean Commission under contract METANET4U (Grant Agreement 270893). Authors also gratefully acknowledge the contribution of the Valletta Local Council.

1 Introduction

The Freedom of Information Act 2008 is a milestone in so far as citizens' rights are concerned: it enables the public, directly or indirectly (through investigative journalists), to disclose that information which the public authorities have not rendered public. In this respect, a freedom of information law brings with it more transparency on the working of the public administration rendering it more accountable to different audiences, including not just the public at large, but also to other, more specialized levels of analysis. But although, in principle, governmental resources are freely available, in practice they are not that easy to obtain. Nor can they be browsed over the internet, be queried, or subjected to automated annotation techniques.

This paper is presented as a result of the unexpected availability of a archive of language data originating from various Maltese Local Council. Malta has 68 Local Councils – 54 in Malta and 14 in Gozo. Local Councils are regulated by the the Local Councils Act, modelled on the European Charter of Local Self-Government (Council of Europe), according to which a Local Council "shall be a statutory local government authority having a distinct legal personality and capable of entering into contracts, of suing and being sued, and of doing all such things and entering into such transactions as are incidental or conducive to the exercise and performance of its functions as are allowed under the Act."

As a consequence of this status, the documents flowing through a Local Council are quite diverse in several different senses, namely:

- **Genre.** The data consists of minutes, memos and data. Each of tese have their own special style.
- **Subject Matter.** The data indicates a wide range of topics ranging from road repairs to social services.
- **Language.** The data is mostly monolingual but is in two languages - both English and Maltese. Not all Maltese text is correctly written (e.g. omitting use of the normal characters instead of Maltese ones)
- **File Formats.** Files do not have uniform structure. That is, there is no overall priniciple of organisation. Particular kinds of document are not written in a uniform style. Minutes, for example, vary according to author. File formats are predominantly word and excel.

1.1 Motivation

A superficial perusal of the dataset suggests that official publications are important in the sense of offering an immense range of opportunities that could benefit three categories of usage:

- Research and development in language technology. The data contained provides examples of Maltese officialese. This is of interest from the stylistic and lexical point of view and could in principle exploited by various writing aids
- Research in the Humanities and Social Sciences. As we shall see, the files contain information that would be useful for the purposes of historical or sociological research concerning the different localities.
- Decision Support within Local Council. In Malta the Local Council system is such that information contained in archives has a tendency to be forgotten. Often, effort that has been expended discussing a particular problem is repeated when, years later, the same problem is rediscussed by a new set of Council members. Another issue is that the same problems tend to crop up in different Local Councils. Solutions that have been developed in one Council ought in principle to be available for scrutiny by another. Under the present arrangements, it is difficult to achieve this level of transparency. Although not explicitly addressed in this paper, we would like to see the introduction of a browsing system for official documents which would improve access to content within the Council system.

The remainder of this document explores the possibilities for improving access to these materials and is structured as follows. Section 2 gives a rough indication of the corpus content. Section 3 outlines the achievements so far and challenges ahead and section 4 provides a set of objectives. We conclude in section 5 .

2 Corpus Content

The corpus comprises around 6,753 files, containing 13 different types of file formats (including a type for the files saved in an undefined format). The size of the dataset is 1.15GB. We have not yet determined the number of words it contains.

This collection is made up of different governmental resources, organized in two main sections: Minutes and Memos. The next two sections give an overview of their content.

2.1 Minutes

The minutes collection, forming 93% of the obtained corpus, is further organised by year, covering the years from 2007 till 2010. Each year is then categorised by locality, thus having 68 sub-collections (with the exception of the year 2007-2008, containing only data of 59 localities) of information. Each sub-collection (representing a local-council) contains a number of:

- Word documents – listing all the events that took place during the meeting.
- Excel documents – listing all the financial data, during a specific period of the year.

The information embedded in the word documents, includes maintenance issues, upcoming event details, obstacles that the locality is facing, and more. Additionally, these documents also contain information about how such obstacles are overcome, maybe also based on previous solutions to similar problems.

After briefly evaluating the collection, we noted that there seems to be no uniform structure amongst the localities, resulting with a different document structure for each local council. Having said this, each local council seems to retain the same document structure over the years.

A more troublesome observation concerns the inconsistent usage of Maltese characters. Not all documents are written using the correct Maltese characters. Furthermore, many documents are mixed in the sense that they may also contain English text embedded with Maltese.

2.2 Memos

The memos collection is much smaller than the minutes, adding up to 7% of the whole corpus. The collection covers the years from 2008 up to 2011.

However, in contrast to the minutes' collection, the files are only categorized by year, holding different Governmental memoranda which were made available during that particular year.

3 Aims and Objectives

Document collections of this kind are probably extremely common, but at the moment only rarely

accessed. Even those who are allowed to access them have difficulty finding the information that they contatin. Our primary aim is therefore to provide progressively more sophisticated access to the information contained in the collection by adopting appropriate technical means. We propose the tackle the problem bottom up: from the basic data, through the contents, towards a coherent structure for the collection as a whole.

This aim leads us to the following objectives (in rough order of difficulty)

- Automated Data Normalization: the need to employ standard representations for text and tabular data and to employ automated methods to translate the sources into such representations. This process is not so very different to those employed for the preparation of other corpora, and we intend to reuse machinery that has already been employed in the development of the MLRS corpus (Borg-et-al, 2011) for this purpose, as reported further in section 4.1.
- Automated Data Analysis: the need to extract meaningful data from the collection. Techniques that are clearly relevant include named entity recognition and topic analysis although it is unclear at this stage how well currently available systems will cope with a collection of this kind.
- Automatic Data Organization: access to data would be greatly facilitated if there were some standard methods for structuring the data automatically. The variety of methods actually employed is bewildering, requiring special (i.e. manual) procedures to access information for practically every combination of locality, topic and document type. So a key issue is whether it is possible to devise a suitable classification scheme for bureaucratic documents. Any such scheme has to strike a delicate balance between generality (being able to accommodate a very wide range of document subject matter) and specificity (implementing principles of organization that will actually make a difference to the the retrieval of useful information.

4 Achievements and Challenges

4.1 Data Normalisation

Given that the corpus is composed of documents in different 13 file formats, there are a number of challenges as regards normalisation. Table 1 shows the distribution of file types. Luckily, the bulk of the corpus consists of pdf files (37%), doc files (34%) and xls files (22%) and these have been successfully converted, using mostly automated techniques, into formats that can be further manipulated: doc and pdf files were converted to txt files, while xls files were converted to csv files as shown in the right hand columns of the table.

Source Type	No. files	Target Type	No. Files
pdf	2559	txt	2101
doc	2308	txt	2295
xls	1520	csv	1275
jpg	196		
.rtf	43		
bmp	12		
xlsx	10		
unknown	9		
gif	5		
docx	4	txt	4
tif	3		
zip	2		
htm	2		

Table 1: structure of collection by file type

A challenge for the text files is that most of them do not use the standard set of Maltese characters, and we are experimenting with automatic spelling correction to overcome this problem.

4.2 Data Analysis

The information residing in these files is clearly valuable but not easily accessible. We are proposing to exploit existing well-established information extraction techniques (see Cunningham-et-al 2005) involving for instance named entity recognition and topic analysis to identify key elements of well known document types and to build gazetteers that include the names of people, organisations, places and quantities. For example, our collection contains a large number of meeting minutes and within documents of this type we would propose to identify

- Which councillors attended the meeting
- The agenda proposed, and decided upon for the meeting
- The different topics, or issues which were to be covered during the meeting
- Other relevant information (e.g. decisions reached) in concordance with the meeting.

We would expect to find other key properties for other kinds of documents.

4.3 Data Organisation

The data is currently organised in a very rudimentary way. Furthermore there are inconsistencies in the way data has been organised by the several Local Councils that have contributed to the collection. We are therefore proposing a kind of intelligent cataloguing system based on sound principles of organisation. We believe that there are certain similarities between on the one hand, the problem of structuring document collections of the kind described and on the other, the organisation of repositories for linguistic resources in general. Techniques which apply to the second problem might fruitfully be applied to first one. Consequently, we will base our solution to the organisation of data around three major components:

- Definition of different document types together with their respective structuring principles. Here we envisage to approach the problem of structuring bureaucratic documents by developing a system of metadata categories not entirely dissimilar to the system already developed for the description of linguistic resources with the METANET4U project (METANET4U deliverable D4.1). This system would then serve as a skeleton into which the actual documents could be fitted.
- A system for intelligently mapping document resources into the document catalogue.
- A system for browsing the contents of the collection according to different principles of organisation such as locality, date, topic.

This system might also have the potential to offer a suggestion facility whereby solutions to common problems and frequently asked questions might be pooled in order to improve local decision making

5 Conclusion

The starting point for this paper was a document collection of a kind which is extremely common, extremely diverse and whose contents could be better exploited. We have described an approach which has the potential to transform a passive document collection only accessible to a few into an information rich resource available to many through the use of mainly existing technologies.

6 References

Borg, C., Fabri, R., Gatt, A., Rosner, M., Maltese and the Digital Age: Developing Electronic Language Resources for the Maltese Language, Linguistics Circle Presentation, University of Malta, 2011

Cunningham, H. Information Extraction, Automatic, Encyclopaedia of Language and Linguistics, Elsevier, 2005

From recording to retrieving: a proposal of a complete system for semi-automatic reporting for local and national governments

Maria Palmerini[*], Ruben Cerolini[*], Giulio Santini[*], Francesco Cutugno[^]

[*] Cedat85, Roma; [^] LUSI-Lab@University of Naples - Italy

{m.palmerini, r.cerolini, g.santini}@cedat85.com, cutugno@unina.it

Abstract

The system we present here gives the possibility of bringing multimediality into the process of information retrieval from audio, video and Italian texts derived by parliament reports. The aim is not only to improve and increase the different ways the official documents can be watched and listened to and retrieved, but also to let all this information be available for different categories of users. Cedat 85 has produced a web service thought to satisfy the requests of the Basilicata Region Council and Verona Town Council, but that, given its premises, aims to be applicable in a wider range of parliamentary environments.

Keywords: multimedial indexing, automatic speech recognition, web service

1. Introduction

In Italy, official reports of the local and national governments involve different activities such as audio recording, video recording and written reporting: these activities are presently often run separately. Cedat 85 has a long experience in the field of official transcriptions and reporting. One of its main qualities is to have both the capability of providing reporting services, and developing new technological solutions in the field of speech processing and multilevel work-flows management.

The system we present here gives the possibility of bringing multimediality into the process of information retrieval from audio, video and Italian texts derived by parliament reports. The aim is not only to improve and increase the different ways the official documents can be watched and listened to and retrieved, but also to let all this information be available for different categories of users, including hearing-impaired subjects and non native people.

2. Multimedial fruition of parliament proceedings

2.1 A Web Portal

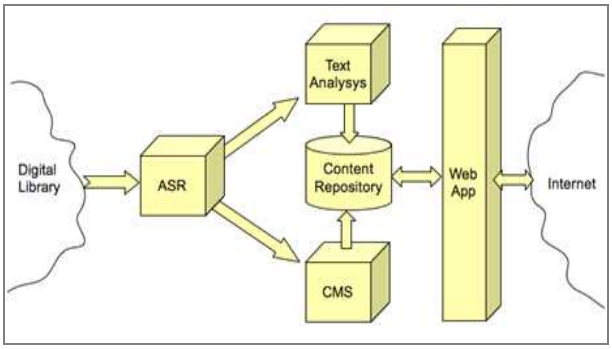
Considering the two National Chambers (Parlamento and Senato), the 20 Regions and the thousands of town halls in which public speeches and other types of formal proceedings are daily produced and archived, Italy, not differently from most countries, requires a reliable system for storing and giving access to the huge amount of data that is generated and that is spread over the whole country. Unfortunately no specific standard has yet been proposed to collect, integrate – when possible – and make this data available to the public.

Cedat 85 has been involved in this field and has produced an initial proposal thought, in principle, to satisfy the requests of the Basilicata Region Council, but that, given its premises, aims to be applicable in a wider range of parliamentary environments. The preliminary but already operating and convincing product of this work can be studied by visiting:

http://www.piuvoce.it/resocontazione_multimediale/

Subsequently Cedat 85 started a collaboration with LUSI-Lab (Language Understanding and Speech Interface Laboratory) at the University of Naples "Federico II" with the aim of moving toward a more robust standardization of the formats and retrieval procedure used in the realized application.

2.2 System architecture



The key element of the system is an Automatic Speech Recognizer (ASR) engine operating on the speech recorded during the parliament sessions. Cedat 85 system from the beginning was trained with a pretty large amount of data, as the aim was to use it to decode spontaneous speech of different environments.

The first model that was created for Italian language was the political one; the acoustic model was trained on 200 hours of speech and the vocabulary, made of 100K words, and was derived by a language model trained on texts for more than 2 millions of words.

In further adaptations to some other environments (media, justice, telephony), the acoustic model has been properly trained and the vocabulary has been increased up to 200K words with exceptional text coverage.

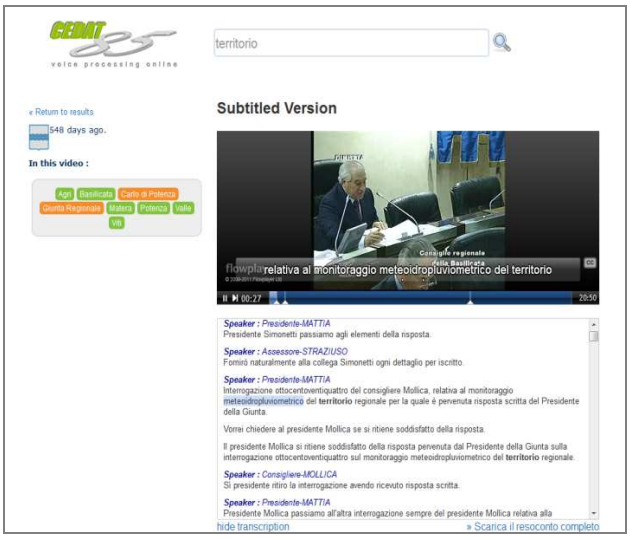
In the application we are presenting here, the temporal progression of each event in the session is strictly related to the audio file and it is transformed into a sequences of words whose start and end time are systematically recorded into an XML file as they are retrieved by the Cedat 85 ASR engine. The produced XML file specifically contains a time-code label for each word given in output by the ASR. This makes possible to associate any word to the time interval in which it is uttered. In the next picture an example of how the XML code appears is given.

```
−<annotated-call xml:lang="it-IT">
  −<call-data channels="2" model="03">
      <callid>0MM0-1</callid>
      <call-back-uri>ftp://10.0.0.70:8181</call-back-uri>
      <agentid channel="0">1</agentid>
      <url>ftp://10.0.0.70:8080/0000-0013-03-0MM0-1.wav</url>
    </call-data>
  −<annotation>
    −<type id="transcription" nbest="1" reading="0">
      +<sentence end="24660" lang="ita" speakerID="0000" start="2785"></sentence>
      −<sentence end="28920" lang="ita" speakerID="0001" start="24990">
          <item end="25500" start="25270">con</item>
          <item end="25920" start="25500">Cai</item>
          <item end="26600" start="26020">intendiamo</item>
          <item end="27300" start="26600">centrale</item>
          <item end="27820" start="27340">allarme</item>
          <item end="28620" start="27820">interbancaria</item>
        </sentence>
      −<sentence end="32180" lang="ita" speakerID="0001" start="29080">
          <item end="29580" start="29370">con</item>
          <item end="29850" start="29580">sette</item>
          <item end="30340" start="29850">zero</item>
          <item end="30529" start="30370">la</item>
          <item end="31190" start="30530">procedura</item>
          <item end="31960" start="31220">anagrafiche</item>
        </sentence>
      +<sentence end="45840" lang="ita" speakerID="0001" start="32410"></sentence>
      </type>
    </annotation>
  </annotated-call>
```

The video signal, usually taken in the chamber and showing the speaker and a part of the remaining people in the meeting, is then temporally indexed using the word sequence as time anchors. In some cases the ASR is reviewed by a human text editor especially when Named Entities and turn taking become problematic.

Audio, text resulting from ASR operations and video, are then aligned to generate the multimedial archive delivered through the portal.

Text is indexed by means of a Google-like search engine that makes it easy to browse all documents. The used engine is architecturally composed by different modules both proprietarily developed by Cedat 85 and recurring to third parties tools as well. The retrieve process, which was started with a string search, produces the appearance of all the video portions containing a related audio signal where the searched word is uttered, the whole text containing the target is shown and the user can point and click on each word in the text causing the audio and the video to skip to that position.

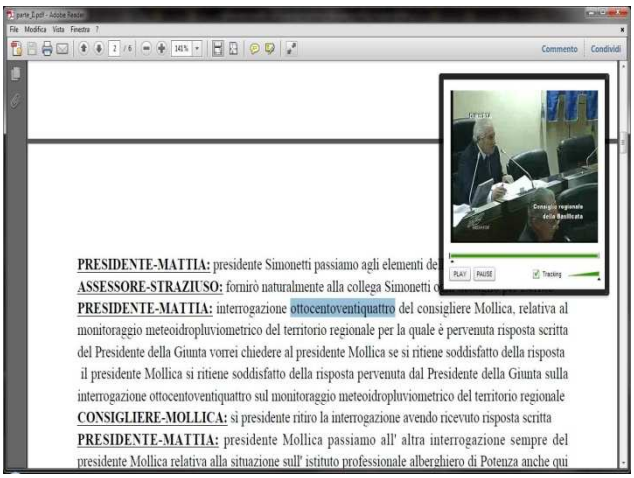


2.3 The multimedia pdf

Portal users can download and consult on their computers as many data excerpts as they want. The system allows them to audio and/or video edit any preceding portion they selected during retrieval process and save the resulting document in a file using the multimedia PDF format.

This format allows the alignment of both the audio and the video to the textual transcription without losing the usual pdf interface and functions, such as the porting toward any environment, printing and exporting.

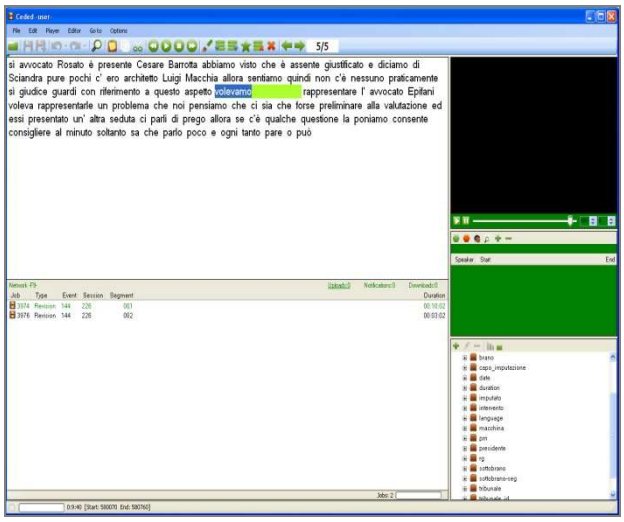
Also in this case, as in the web application, the alignment between audio (and video) and text allows a point & click on each word of the text in order to listen to the audio. Moreover, a tracking function highlights the words spoken in the audio(-video).



3. Written and spoken language in official documents: problems and possible solutions

Finding the features that express the difference between spoken and written language and using this knowledge to increase the quality of transcription work in general – not just in official contexts – is indeed one of the main issues in the field of speech-to-document activity. As linguists know well, these two varieties of language diverge not

only in the channel they use (respectively acoustic and visual), but also (and consequently) because they follow very different syntactical and textual rules.
The consequence is that the activity of transferring a text from the spoken form (used to produce it) to the written form (needed to receive it, and furthermore requiring conformity to specific rules if the text has to be used in particular environment) needs some changes and adjustments. To mention an example, in many cases, repetitions, hesitations and disfluencies in general have to be removed from the written text, to make it readable; or a wrong name will have to be corrected. Such modifications to the text can generate a difference between the final text and the time aligned xml, causing problems in the automatic indexing and alignment among audio, video and text.
In our system a back-end module for text editing has been thought to reduce this kind of problem, it is used to modify the xml file directly, in order to save the time labels of each token and the original alignment after some corrections are made.



As can be seen in the screenshot, the operator interface shows only the text of the document, in order to make the editing as easy as using a common and familiar text editor; but each change on the text will produce a change in the xml file as well.
This module can only reduce and not completely solve the problems that can be found when transferring spoken language into a written text. For example, it's still not possible to modify the order of single words or of strings of text, without causing a mismatch in alignment. It's clear that if we want to maintain a word-by-word audio-text alignment, their order can't be changed and this can generate a certain difficulty in reading the text (that is always structured as a spoken text). Nevertheless, the entire system we propose has been projected to be used as a multimedia database, where the user can not only read or watch or listen to the data separately, but where these different channels are completely integrated and can all be used together at the same time, with the aim to extend and increase the ways that information can be retrieved.

4. Conclusions and further developments

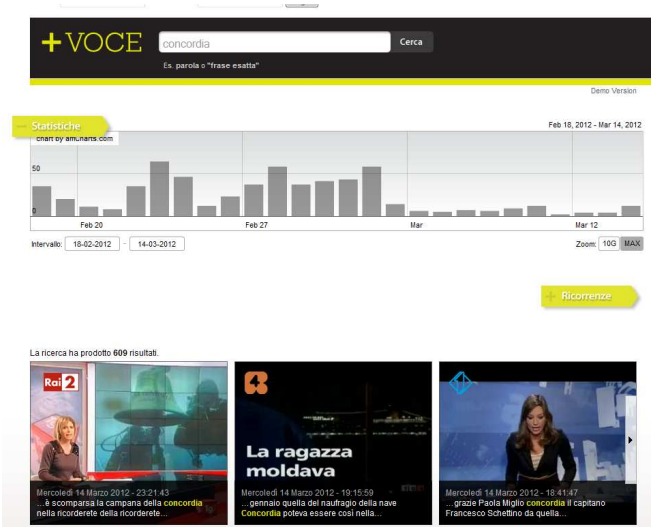
The presented system in this paper is already operating for some local governments and is not a prototype. Two different applications of the same engine devoted to two different customers can be encountered here:

[*http://www.piuvoce.it/resocontazione_multimediale/*

and

http://demo.piuvocepa.it/ccverona_premium/?language=all]

Moreover, the system is already tailored to respond to the requirements of a wider range of institutional environments, such as the national Parliament. As far as we (and our customer) know, this is the only project of this kind in Italy.
The system is presently in continuous evolution, in order to improve and enrich it with new features.
At the moment, the implementations we are working on are in the field of speech analytics and knowledge management. Besides, a module for captioning could be implemented, allowing a better and wider accessibility of the official reports.
In Italy there are some regions (such as Trentino Alto Adige or Valle d'Aosta) where there is more than one official language, also for official documents. For this kind of environments, it would be interesting to add also other languages in the ASR system, in order to make the system multilingual and, again, improve the accessibility to the documents.
Besides the applications to institutional environments, the system we present here is also provided, with some re-planning of the main interface, to be used in other environments.
First of all, we would like to mention another application by Cedat 85 named Piuvoce [*http://www.piuvoce.it*]. Piuvoce is a media monitoring application, where a very large amount of broadcasts form the main national tv and radio channels is recorded, automatically transcribed and indexed so that each single word can be retrieved in the whole archive. Besides, the application contains an editing module where the user can select the piece of transcription of his interest and cut that piece of both video and transcription. All the video clips can be saved in the personal area and they can be appended one to the other in the order chosen by the user. This way it's possible for the user to collect all his cuts and create his own video review about his favorite subjects.
This service can be further enriched: it's possible for the user to receive an alert via sms each time that a given word or entity appears in the news, with a minimal delay. In the screenshot we show the statistics page about the word “Concordia”, which is the name of a very big boot that sank in February 2012 right in front of Tuscan coasts.



Other possible applications of the system we present are in general all the events where an audio (and video) is produced and needs to be saved to be available to a group of users. We can mention:

- conferences: let's think to medical conferences, where a video of a surgery operation is showed and discussed;
- education: university classes could be audio and video recorded, automatically transcribed and indexed, and then be available to students, including hearing-impaired and foreign ones.

Automatic multilingual annotation of EU legislation with Eurovoc descriptors

Vidas Daudaravičius

Vytautas Magnus University
Vileikos 8, Lithuania
v.daudaravicius@if.vdu.lt

Abstract

Automatic document annotation from a controlled conceptual thesaurus is useful for establishing precise links between similar documents. This study presents a language independent document annotation system based on features derived from a collocation segmentation method. Using the multilingual conceptual thesaurus EuroVoc, we evaluate the method, comparing it against other language independent methods based on single words and bigrams. Testing the method against the manually tagged multilingual corpus Acquis Communautaire 3.0 (AC) using all descriptors found there, we attain improvements in keyword assignment precision from 50.7 to 57.6 percent over three diverse languages (English, Lithuanian and Finnish) tested. We found high correlation between automatic assignment precision against document length and language features such as inflectiveness and compounding.

1. Introduction

Automatic multilingual document annotation requires the understanding of differences among the languages and features that are useful for the classification. In this study we take three diverse languages such as English, Lithuanian and Finnish as these languages differ in inflectiveness and compounding aspects. Finnish language is highly inflected and compounded, Lithuanian is highly inflected and no compounds undertakes, and English is not neither inflected nor compounded compare to Lithuanian and Finnish. The main goal of out study is to compare different methods for building feature vectors and to understand how quality of automatic annotation depends on the document size. In the Acquis Communautaire (AC) corpus (Steinberger et al., 2006) we find various data in different languages.

2. Related work

The EuroVoc thesaurus and all the legal documents indexed with its categories have been widely used in natural language processing tasks. (**?**) created one of the most widely used multilingual aligned parallel corpora, the JRC-Acquis, in 22 languages from manually labeled documents. Automatic Eurovoc indexing is a multi-label classification task, successfully applied and discussed abundantly in literature (e.g. (Manning et al., 2008)). Here, we focus on Eurovoc-related work. (Mencía and Fürnkranz, 2010) applied three different multi-label classification algorithms to the Eur-Lex database of the European Union indexed with Eurovoc addressing the problem of memory storage for classifiers. In our work, we are interested in investigating if collocation segmentation for the document representation can improve precision with minimal increase in computational costs for (non-)inflected and (non-)compounded languages such as English, Lithuanian and Finnish. Our approach is based on the work proposed by (Pouliquen et al., 2003) and highly related to (Daudaravicius, 2010).

3. Eurovoc descriptor assignment

3.1. Data Set

Experiments were carried out on the Acquis Communautaire (AC) corpus (Steinberger et al., 2006). This large text collection contains documents selected from the European Union (EU) legislation in all the official EU languages. Most of these documents have been manually classified according to the EuroVoc thesaurus (of the European Communities, 1995). In this study we take only those documents that have assigned EuroVoc descriptors. The AC documents are XML marked. The body and signature parts of the documents were used in our study. Annexes were excluded.

The corpus was split into the development part and test part. For each language 95 percent of documents were randomly selected for the development corpus and 5 percent of the documents (about 500 documents) were used for the test corpus. There is no overlap between the development corpus and the test corpus. The documents for development and test corpora for different languages are not the same as AC corpus is not fully parallel. Thus, the results for different languages can be compared with caution. Nevertheless, the test corpus of each language is the same for each experiment. The number of descriptors depends on the length of a document. The maximum number of descriptors for very short documents is very low compare to documents of other lengths, and is four times bigger than the average number of descriptors. And the average number of descriptors for documents of the length between 251 and 1000 words is lower than of other length documents (see Table 1). This dependency of average number of descriptors can be related to the different types of the documents that are included in AC corpus. Different types of documents could contain a different amount of context necessary to understand what a document is about. It might be that the reason for the higher number of descriptors for short documents is to use descriptors as an additional or contextual information useful in query systems.

No requirements were applied for descriptors to occur in several documents. The average number of descriptors per document in test corpus was 5.4. As we can see in Figure 1, the largest part of documents are assigned to 4, 5 or 6 descriptors and only several to more than 10 or less than 2 descriptors. The ranges of document lengths were set to 6 intervals as shown in Table 1. In Table1 we see that the

ID	Document length		Number of documents			Number of descriptors per document (En)	
	From	To	En	Lt	Fi	Maximum	Average
1	1	100	11	37	23	11	5.6
2	101	250	790	1409	1706	18	5.6
3	251	500	4970	5852	6719	20	5.2
4	501	1000	7065	6263	5647	17	5.2
5	1001	2500	3853	3522	3430	24	5.7
6	2501	...	3684	3152	2892	20	5.9

Table 1: Corpus data by document length

most of the documents are in the range of document length between 250 and 500 words and there is a slight move of the distribution of document length by language. The median of the document length for English is 583 words, for Lithuanian is 435 words and for Finnish is 396 words. This move of median shows that the document length is influenced by the inflectiveness and the compoundness. The highest influence for reduction of the document length have inflections and then compounds.

Before we trained and tested our classifier, the AC corpus underwent a basic preprocessing consisting of lowercasing and tokenisation. The numbers were left as is. We decided not to apply any language-dependent preprocessing, such as lemmatisation, since our study is intended to work in a multilingual environment.

3.2. Collocation Segmentation

Collocation segment is a sequence of words or terms in a text and the boundaries of this segments depends on the surrounding context and does not depend on the number of its occurrences in a corpus. This definition differs from collocation in terms that collocation is based on the number of occurrence in a corpus. Collocation is defined as is 2 or more word length sequence while collocation segment can be of any length (even single word) and this length is not defined in advanced. The collocation segmentation method differs from other widely used statistical methods for collocation extraction that are mainly dictionary based methods using frequency filtering (Tjong-Kim-Sang and S., 2000; Choueka, 1988; Smadja, 1993) or syntactic rules (Lin, 1998).

For the collocation segmentation the Dice score is used to measure the association strength of two words. This score is used, for instance, in the collocation compiler XTract (Smadja, 1993) and in the lexicon extraction system Champollion (Smadja et al., 1996). Dice is defined as follows:

$$Dice(x_{i-1}; x_i) = \frac{2 \cdot f(x_{i-1}; x_i)}{f(x_{i-1}) + f(x_i)}$$

$f(x_{i-1}; x_i)$ being the number of co-occurrence of x_{i-1} and x_i, and $f(x_{i-1})$ and $f(x_i)$ are the numbers of occurrence of x_{i-1} and x_i in the training corpus. If x_{i-1} and x_i tend to occur in conjunction, their Dice score will be high. There are many other association measures such as Mutual Information (MI), T-score, Log-Likelihood and etc. A detail overview of associativity measures could be found in (Pecina, 2010). MI and Dice scores are almost similar in the sense of distribution of values (Daudaravicius and Marcinkeviciene, 2004). We decided to use Dice score for the reason that MI and Dice are similar by distribution (see (Daudaravicius and Marcinkeviciene, 2004)) but the range of Dice values are between 0 and 1, and Mi range of values depends on a corpus size. Other measures like T-score could be used also because this measure produces different segmentation (see (Henríquez Q. et al., 2010)), but this measure requires much more calculations and extends processing time, and the range of combinability values depends on a corpus size also.

The associativity values are used to produce a discrete signal of a text as in (Daudaravicius, 2010) study. A text is seen as a changing curve of Dice values between two adjacent words (see Figure 2). This associativity value curve is used for detecting the boundaries of collocation segments. The boundaries are set as following.

3.3. Setting segment boundaries with Threshold

The boundary is set between two adjacent words in a text where the Dice value is lower than a threshold. In (Daudaravicius, 2010) the threshold value is set manually and is absolute value for whole segmentation. To explore different levels of threshold we introduce dynamic threshold which defines the range between the minimal and the average associativity values of a sentence. 0 equals to minimal associativity value and 100 equals to the average value of the sentence. Thus, the threshold value is expressed as percents between the minimal and the average associativity values. If the threshold is set to 0 then this means no threshold filtering is used at all and no collocation segment boundaries are set using threshold. The main purpose of using threshold is to keep only 'strongly' connected tokens. On the other hand, there is a possibility to set the threshold at the maximum value of associativity values. This would make no words combined into more than single word segments, i.e. collocation segmentation would be equal to simple tokenisation. Threshold gives the possibility to move from single word tokens to whole sentence tokens by changing threshold from minimum to maximum value of the sentence. In our study we explore two threshold level that are 0 percent (Seg 0) and 50 percent (Seg 50) to understand the influence of threshold to classification results.

In Figure 2 we can see that the average threshold for Lithuanian is only 40 percent as high as that for the other two languages. If one or several strongly collocated words oc-

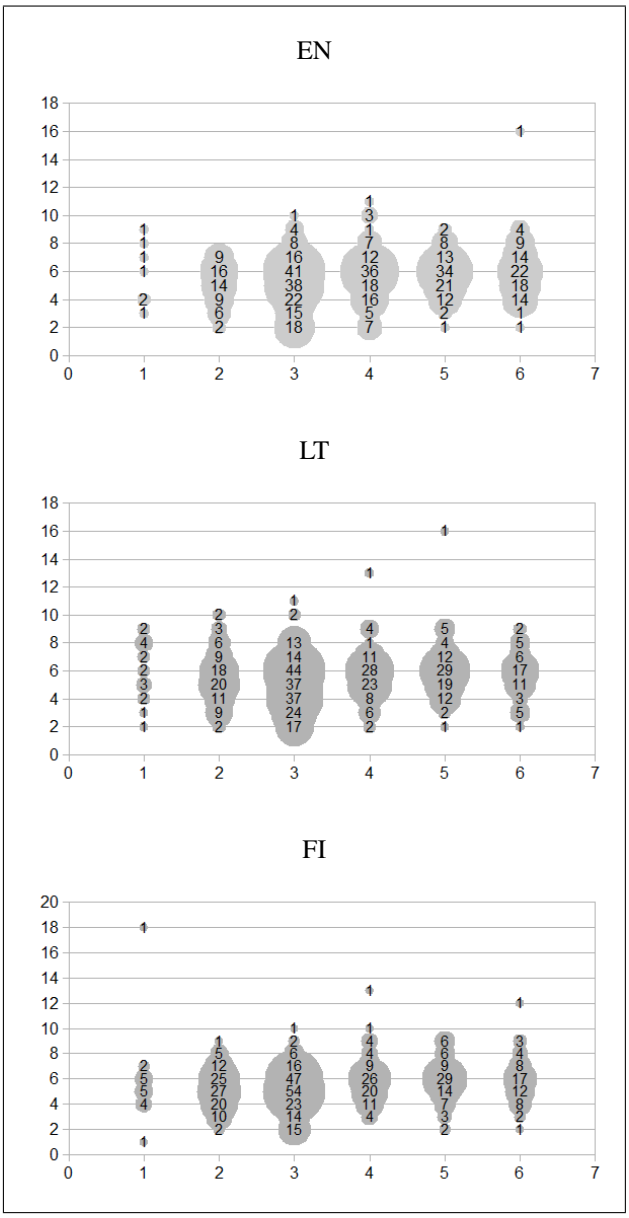


Figure 1: The number of test documents (Bubble size) against document size (X-axis) and the number of manually assigned descriptors (Y-axis).

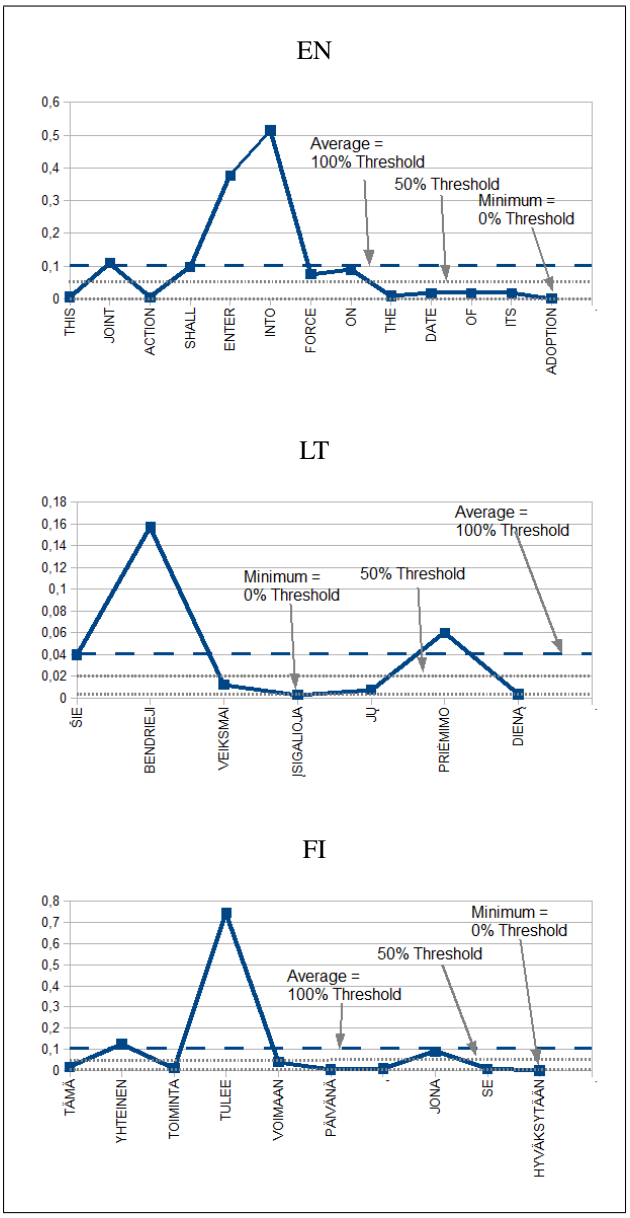


Figure 2: Combinability curve of the same sentence in three explored languages. (Y-axis represents Dice score)

cur in a sentence then the average of collocability values goes up quickly. By the accident Finnish example contains one strongly collocated word pair. On the other hand, most of combinability values are between 0 and 0.1. Finnish and Lithuanian are highly inflected and use less prepositions and no articles. Articles and prepositions tend to have higher combinability value in word pairs and this gives up the average value of combinability. By setting dynamic threshold level we make higher restrictions on generating long collocation segments. Thus, dynamic threshold level gives flexible means to control the length of collocation segments.

3.4. Setting segment boundaries with Average Minimum Law

The average minimum law (AML) is introduced in (Daudaravicius, 2010). AML is applied to the three adjacent associativity values and is expressed as follows:

$$\frac{Dice(x_{i-3}; x_{i-2}) + Dice(x_{i-1}; x_i)}{i} < Dice(x_{i-2}; x_{i-1})$$

$$\Rightarrow set\ boundary(x_{i-2}, x_{i-1})$$

The boundary between two adjacent words in the text is set where Dice value is lower than the average of preceding and the following Dice values. AML defines collocation segment boundaries by the context of associativity values and is not related to frequencies of segments nor threshold. In Table 2 we can see all collocation segments that contain substring *language* and that were produced by the collocation segmentation with threshold level 0. We can see that

Segment	$f(x)$	Segment	$f(x)$	Segment	$f(x)$	Segment	$f(x)$
languages	661	language technologies	5	language requirements	2	language classes	1
official languages	539	language training for	5	language resources	2	language combination	1
and norwegian languages	380	languages accepted	5	language training and	2	language competency	1
official language	343	languages used	5	language unit	2	language configurations	1
authentic language	134	minority language	5	language would	2	language course	1
language versions	116	specifies language	5	languages :	2	language danish	1
language learning	83	union languages	5	languages dropped	2	language department	1
swedish languages	77	a language	4	languages portal	2	language dependent	1
original language	76	georgian languages	4	languages spoken	2	language division	1
language courses	74	guilanguage	4	languages spoken inside	2	language education	1
english language	58	in different languages	4	linguistic aspectslanguage	2	language of proceedings	1
language version	52	intelligible language	4	migrant languages	2	language of transmission	1
working languages	46	language combinations	4	minority language periodical	2	language optical	1
language skills	43	language portfolio	4	multiple languages	2	language other	1
or languages	38	language testing	4	native languages	2	language paves	1
community languages	37	language training ,	4	official community languages	2	language peculiarities	1
language or languages	36	languages of	4	original language version	2	language preferred	1
the language	35	languages permitted	4	prescribed languages	2	language preparationpreparatory	1
spanish languages	33	local language	4	serbian languages	2	language publications	1
foreign languages	30	of language	4	slanguageiso	2	language requirements (	1
10 languages	29	passive languages	4	slovak languages	2	language rules	1
norwegian language	27	per language	4	source language	2	language searchability	1
slovenian languages	25	regional languages	4	spoken languages	2	language section	1
working language	25	that language	4	swedish language	2	language selected	1
language books	24	these languages	4	teaching languages	2	language shortages	1
italian language	23	accessible language	3	- language	1	language tends to	1
language accepted	23	accommodate language	3	- language secretaries	1	language testing	1
language indicated	20	and language	3	, language and	1	language training and cultural	1
languages ,	20	bulgarian languages	3	? which language	1	language used .	1
15 languages	18	danish language	3	language industries	1	language used throughout	1
another language	17	greek language	3	) other languages	1	language verion	1
authentic languages	17	in several languages	3	11 languages	1	language via	1
french language	17	language and	3	active languages	1	languages .	1
language field	16	language facilities	3	active languages are	1	languages . besides	1
foreign language	15	language integrated learning	3	agreed languages	1	languages accepted for	1
minority languages	15	language needs	3	and active languages	1	languages and form	1
languages and	14	language problems	3	and working language	1	languages are equal	1
, language	13	language professions	3	appropriate language	1	languages can flourish	1
language competence	13	language understood	3	belarusian language	1	languages different	1
language arrangements	12	languages acceptable	3	bosnian languages	1	languages is	1
language training	12	languages allowed	3	catalan language	1	languages of our	1
language proficiency	11	learn languages	3	clearer language	1	languages prescribed	1
language units	11	local language requirements	3	common language	1	languages requested	1
language works	11	maltese language	3	concepts and language	1	languages were	1
language acceptable	10	national language	3	correct language	1	local languages and	1
language or	10	official community language	3	definite language	1	moldavian languages	1
language customary	9	other languages	3	description language as	1	moldovan languages	1
language knowledge	9	russian language	3	determining the language	1	more languages	1
language of	9	technical language	3	dutch language	1	ms language	1
language teacher	9	turkish language	3	emotional language	1	ms languages	1
procedural languages	9	language	2	everyday language	1	new language	1
slovene languages	9	21 languages	2	finnish language	1	new languages	1
used languages	9	a second language	2	foreign language competence	1	norwegian languages	1
community language	8	additional languages	2	foreign language teacher	1	of languages only	1
several languages	8	albanian languages	2	foreign language teaching	1	of several languages	1
and working languages	7	clear language	2	foreign languagemodern	1	other language	1
different language	7	five languages	2	format and language	1	overcome language	1
language barriers	7	german language	2	french - language	1	pages per language	1
language proficiency rating	7	in another language	2	further languages	1	per language unit	1
language teachers	7	inclusive language	2	impose language requirements	1	pose language	1
language teaching	7	interpretation into that language	2	in more languages	1	romanian language	1
romanian languages	7	interpretation language profile	2	including language	1	schengen languages	1
russian languages	7	interpreters per language booth	2	indigenous languages	1	sign language	1
second language	7	language (	2	intended language	1	some languages	1
different languages	6	language books abroad	2	interoperable language	1	special language	1
host language	6	language certification	2	involve language expertise	1	stronger environmental language	1
language endorsements	6	language competences	2	italian languages	1	target languages requested	1
language used	6	language comprehensible	2	japanese languages	1	teach languages	1
languages in	6	language employed	2	korean language	1	technological language	1
of languages	6	language endorsement	2	kurdish language	1	the languages	1
20 languages	5	language error	2	language ,	1	those languages	1
eu languages	5	language is	2	language acquisition	1	understood language	1
irish language	5	language issues	2	language areas	1	using different languages	1
language chosen	5	language learners	2	language assistants	1	various language	1
language configuration	5	language preparation	2	language borders	1	which languages	1
language diversity	5	language regimes	2	language broadening	1	xenophobic language in	1
language regime	5						

Table 2: Collocation segments containing substring *language*

most of the segments are two-word length segments and many of them correspond to the noun phrase. Some segments are not grammatical and belongs to other phrases. For instance, *and norvegian languages* is the end of the enumeration of languages, and clearly is not grammatical.

3.5. Experiment settings

The collocation segmentation was performed on each language corpora using two thresholds: threshold 0 (Seg 0) and threshold 50 (Seg 50). We choose to use relatively low segmentation threshold level because higher threshold reduces collocation segments close to single word segments quickly. All descriptors and all words or segments were taken into consideration in order to keep the real environment and to see the real possibilities of automatic Eurovoc descriptor assignment. We did not use any stop word lists to remove common words. All words were kept as is as we did not want to use linguistic resources that are hard to prepare and that are ambiguous and domain dependent. For each experiment we produce feature vector for each test document and feature vectors for each descriptor on the basis of one large meta-text (concatenation of all texts indexed with this descriptor). We explore unigram (uni), bigram (bi), segmentation with threshold 0 (seg 0) and segmentation with threshold 50 (seg 50). We decided to modify widely used TFIDF formula and named our modification as normalized TFIDF with confidence adjustment which is defined as follows:

$$NCTFIDF(x) = \\ = \frac{TF(x) * avg|Doc|}{|Doc|} * ln(\frac{N - D(x) + 1}{D(x) + 1}$$

$TF(x)$ being the raw frequency of token x in the document or descriptor profile, $avg|Doc|$ being the average document length in the corpus or sum of frequencies of descriptor profile items, $|Doc|$ being the length of the inspected document or the sum of frequencies of descriptor profile items, N being total number of documents in the corpus or total number of descriptors, $D(x)$ being the number of documents or profiles the token x occurs. The idea of this modification is to change the order item and to put items, that are frequent and common in documents, at the end of the list and to give them negative values. In our experiments we removed all items with negative values. This worked as stop-word filtering. Normalization allows to reduce the impact of document size when we count the distance between documents and descriptors.

The cosine distance was used to evaluate distance between document and descriptor profile feature vectors. For each document in the test corpus we calculate the cosine distance to each descriptor profile then we sort by cosine similarity and take exact amount of descriptors as it was manually assigned. For instance, if there was manually assigned fourteen descriptors we take the same number of descriptors with highest cosine similarity and evaluate the accuracy of annotation by the intersection of both sets.

We decided to assign the same number of descriptors as it was done manually on the assumption that all manually assigned descriptors are important and the system should follow human behavior. For instance, we took a document of the length between 251 and 500 words from AC corpus which have the highest number of manually assigned descriptors:

> Protocol to the Agreement with Switzerland on the free movement of persons ***
> European Parliament legislative resolution on the proposal for a Council decision on the conclusion, on behalf of the European Community and its Member States, of a Protocol to the Agreement between the European Community and its Member States, of the one part, and the Swiss Confederation, of the other, on the free movement of persons, regarding the participation, as contracting parties, of the Czech Republic, the Republic of Estonia, the Republic of Cyprus, the Republic of Latvia, the Republic of Lithuania, the Republic of Hungary, the Republic of Malta, the Republic of Poland, the Republic of Slovenia and the Slovak Republic, pursuant to their accession to the European Union (12585/2004 COM(2004)0596 C6-0247/2004 2004/0201(AVC))

Full version of this document could be found at http://eur-lex.europa.eu/Notice.do?val=423395
For this document there are 20 manually assigned descriptors:

2901 ratification of an agreement
1633 free movement of persons
4324 Switzerland
12 accession to the European Union
2850 protocol to an agreement
5420 accession to an agreement
1634 free movement of workers
2543 Poland
1255 Hungary
5859 Slovakia
5898 Slovenia
5989 Cyprus
1774 Malta
5619 Estonia
5709 Lithuania
5706 Latvia
5860 Czech Republic
2814 agricultural real estate
5093 acquisition of property
3464 secondary residence

At the moment, our system is not capable to guess the number of descriptors which could fully describe document as human do. In future work our system will be extended to be able to guess the number of descriptors.

4. Evaluation of the experiment results

For the evaluation of experiment results we use precision only as the recall and precision are equal.

The results in Table 3 show that collocation segmentation without threshold applied outperforms other methods

	En	Lt	Fi	Average
Unigram	0.488	0.495	0.540	0.507
Bigram	0.547	0.553	0.535	0.545
Segmentation 0	**0.559**	**0.585**	**0.585**	**0.576**
Segmentation 50	0.505	0.541	0.554	0.533
Number of categories trained	3811	3797	3792	3800

Table 3: Descriptor assignment precision

	En	Lt	Fi
Unigram	208,583	345,292	536,022
Bigram	1,501,113	2,493,255	2,631,438
Segmentation 0	771,974	1,431,953	1,739,846
Segmentation 50	288,481	497,153	748,907

Table 4: Vacabulary size

such as unigram , bigram or collocation segmentation with threshold applied. The same results show that highly inflected and compounded languages such as Finnish can achieve better results compare to less inflected or compounded languages. In (Steinberger et al., 2012) we see that in general there are no significant differences among the results of assigning descriptors for any language. But this can be influenced by the different average document length and the number of stop-words used. For Finnish language it was used only 92 stop-words while for English it was used 1972 stop-words. In (Mohamed et al., 2012) it is explained that stop-word list have big impact for classification. Therefore, we could expect better results for Finnish than for English if the list of stop-words would be longer. Also, the average document length for Finnish is twice longer than for English, 756 and 309 words respectively. Our study show that for all languages we have tested the precision is higher by 3 percents for documents of the length between 251 and 500 words compare to documents of the length between 501 and 1000 words. Therefore, in (Steinberger et al., 2012) we could expect the precision for Finnish to be higher at least by 2 percents than for English.

We compare precision improvement for the different language under the same data using unigrams, bigrams and collocation segments. The results show that for less agglutinative and less inflected language we gain more improvements when collocation segments are used instead of unigrams.

The speed of the indexing task using different methods is related to the size of vocabulary used. In table 4 we can see vocabulary size for different methods. The time used for assigning descriptors is directly dependent on the size of vocabulary. For instance, for our developed system it takes about one minute to calculate the distance between document and all descriptors when unigrams are used. For bigrams it takes about 7 minutes and it is about 2 minutes for collocation segments with 0 threshold. The results show that collocation segmentation allows to improve classification results better than bigrams while indexing time is several times shorter.

4.1. Document size

There is a high correlation between document length and the way the features are used for classification. For documents up to 250 word length the bigram features give worst classification results and outperforms unigram features for larger documents. On average the best results are achieved using collocation segmentation without any threshold applied. In Finnish, bigrams work slightly less well than unigram. This degradation could be related to the size of features each method capture. In Finnish, the dictionary size of bigrams is very long (see Table 4) and it might be that such long dictionary does not help to get good features for classification and the system cannot be trained well.

The manual evaluation of the manual assignment in (Pouliquen et al., 2003) showed that the evaluators working on the English texts judged 74% of the previously manually assigned descriptors as good. This user interagreement is similar to the precision using collocation segmentation for long English documents which is between 61.3% and 87.1%. It is still difficult to judge on the precision of short documents. Table 5 shows the correlation between the document length and the precision: the bigger the document the higher precision we achieve. This could also correlate to the user interagreement: the shorter document the more diverse descriptors are assigned. The more larger the document the more accurate could be the manual descriptor assignment. But this is not discussed in (Pouliquen et al., 2003) and we cannot compare results to make decision whether the size of the document correlates with the user interagreement.

4.2. Short document annotation

The results in table 5 show that short document annotation is hard to solve and it is difficult to get high precision results. This could be easily seen from the short document example (jrcC2006#121#34-en) in test corpus:

> Order of the Court of First Instance of 21 March 2006
> Holcim (France)
> v Commission
> (Case T -86/03) [1]
> (2006/C 121/34)
> Language of the case: French
> The President of the Second Chamber has ordered that the case be removed from the register.
> [1] OJ C 112, 10.5.2003.

For this document there are seven manually assigned descriptors:

5837 action for annulment of an EC decision
215 inter-company cooperation
1549 fine
3263 redemption
1476 interest
5993 cement
4038 EC Commission

Non of the methods captured automatically at least one right descriptor. This example shows that there is not

Document length	En				Lt				Fi				Average			
	Uni	Bi	Seg0	Seg50	Uni	Bi	Seg0	Seg50	Uni	Bi	Seg0	Seg50	Uni	Bi	Seg0	Seg50
0 – 100	0.220	0.170	**0.268**	0.195	0.154	0.154	**0.194**	0.188	**0.269**	0.192	0.261	0.160	0.214	0.172	**0.241**	0.181
101 – 250	0.533	0.523	**0.575**	0.551	0.506	0.458	**0.537**	0.545	0.558	0.519	**0.568**	0.509	0.532	0.500	**0.560**	0.535
251 – 500	0.492	0.511	**0.544**	0.497	0.509	0.574	**0.614**	0.529	0.533	0.508	**0.575**	0.542	0.511	0.531	**0.578**	0.523
501 – 1000	0.467	0.516	**0.540**	0.474	0.459	0.514	**0.527**	0.505	0.504	0.519	**0.522**	0.549	0.477	0.516	**0.530**	0.509
1001 – 2500	0.493	0.591	**0.871**	0.528	0.532	0.624	**0.669**	0.562	0.549	0.582	**0.651**	0.609	0.525	0.575	**0.730**	0.566
2500 – ...	0.497	**0.640**	0.613	0.531	0.552	**0.717**	0.696	0.659	0.657	0.705	**0.732**	0.667	0.569	**0.687**	0.686	0.619

Table 5: Descriptor assignment precision

enough information for the machine to catch good features for classification as it requires wider context understanding.

5. Conclusions

In our study we found that short documents up to 100 word length are hard to annotate and longer ones could be annotated with acceptable precision. For inflected and compounded languages we attain better precision of document annotation with Eurovoc descriptors using unigrams as features.

6. References

Y. Choueka. 1988. Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO Conference on User-Oriented Content-Based Text and Image Handling*, pages 21–24, Cambridge, MA.

V. Daudaravicius and R Marcinkeviciene. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics*, 9(2):321–348.

Vidas Daudaravicius. 2010. The influence of collocation segmentation and top 10 items to keyword assignment performance. In Alexander F. Gelbukh, editor, *CICLing*, volume 6008 of *Lecture Notes in Computer Science*, pages 648–660. Springer.

Carlos A. Henríquez Q., Marta Ruiz Costa-jussà, Vidas Daudaravicius, Rafael E. Banchs, and José B. Mariño. 2010. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 98–102, Uppsala, Sweden, July. Association for Computational Linguistics.

D. Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Eneldo Loza Mencía and Johannes Fürnkranz. 2010. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors, *Semantic Processing of Legal Texts*, volume 6036 of *Lecture Notes in Computer Science*, pages 192–215. Springer.

Ebrahim Mohamed, Maud Ehrmann, Marco Turchi, and Ralf Steinberger, 2012. *Multi-label EuroVoc classification for Easter and Southern EU Languages*. Multilingual processing in Eastern and Southern EU languages – Low-resourced technologies and translation. Cambridge Scholars Publishing, Cambridge, UK.

Commission of the European Communities. 1995. *Eurovoc Thesaurus: Permuted alphabetical version*. Thesaurus Eurovoc. Office for Official Publications of the European Communities.

Pavel Pecina. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation*, 44(1-2):137–158.

Bruno Pouliquen, Ralf Steinberger, and Camelia Ignat. 2003. Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of the Workshop Ontologies and Information Extraction at (EUROLAN'2003)*, pages 9–28.

Frank Smadja, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. 1996. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22:1–38.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19:143–177.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Toma Erjavec, Dan Tufi, and Dniel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC'2006, pages 2142–2147, May.

Ralf Steinberger, Ebrahim Mohamed, and Marco Turchi. 2012. Jrc eurovoc indexer jex - a freely available multilabel categorisation tool. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*.

E. Tjong-Kim-Sang and Buchholz S. 2000. Introduction to the conll-2000 shared task: Chunking. In *Proc. of CoNLL-2000 and LLL-2000*, pages 127–132, Lisbon, Portugal.

GLOSS, an infrastructure for the semantic annotation and mining of documents in the public security domain

Francesca Frontini[1], Carlo Aliprandi[2], Clara Bacciu[3], Roberto Bartolini[1]
Andrea Marchetti[3], Enrico Parenti[4], Fulvio Piccinonno[4], Tiziana Soru[2]

[1]ILC CNR, Pisa, [2]Synthema, Pisa, [3]IIT CNR, Pisa, [4]Genesy, Pisa - Italy
[1]{name.surname}@ilc.cnr.it, [2]{name.surname}@synthema.it, [3]{name.surname}@iit.cnr.it, [4]{name.surname}@genesy.it

Abstract

Efficient access to information is crucial in the work of organizations that require decision taking in emergency situations. This paper gives an outline of GLOSS, an integrated system for the analysis and retrieval of data in the environmental and public security domain. We shall briefly present the GLOSS infrastructure and its use, and how semantic information of various kinds is integrated, annotated and made available to the final users.

Keywords: semantic annotation, text mining, geographic data

1. Introduction

In today's world the access to large amounts of information is crucial in many domains where decisions must be taken in a short period of time or under special circumstances. This is particularly true for the case for official institutions that deal with public security, protecting the human life and environment from natural disasters and other emergencies. Such institutions - such as US *FEMA* or Italian *Protezione Civile* - produce large amounts of public official documents in different kinds of formats (websites, documents, databases, ...) that embody the specialised knowledge produced by their community and that are normally available to everyone within and outside the organization. These documents contain information relating to past events in specific parts of the territory and how they have been dealt with; therefore they can be useful in case of repetition of similar events or if the environmental risk(s) for a certain area are investigated. For instance a document may contain information on a past flooding event in a certain area at a certain date. Such descriptions of events are referred to as *facts*, and can carry relational information about the conceptual structure that is expressed in texts (what happened, when, where, why, who was involved, etc.). Being able to efficiently identify and retrieve facts in a document base is of crucial importance when dealing with emergencies. In particular it would be important for organizations to be able to see all past events that share the same area, or the same cause, or involved the same individuals or groups, or that happened at a given time.

In order to correctly deal with facts it is necessary not only to identify them in texts but also to annotate them in a way that allows the semantic search tools to access them by one of their specific components (location, time, causes, involved entities, ...).

The KYOTO project [1] has achieved the result of building a cross linguistic platform for knowledge sharing that enables an organization to share and to access information. This is performed by providing:

- an environment for linguistic and semantic analysis that enables intelligent text mining and deep semantic search;
- a Wiki environment that allows people in the field to maintain their knowledge and agree on meaning without knowledge engineering skills.

An important idea of KYOTO is that the specialized knowledge which is extracted and defined from the official documents can in turn be fed to the linguistic and semantic analysis chain thus allowing for more and more refined text and fact mining capacities.

The GLOSS (GLObal Semantic System) project [2], building on the experience gained in KYOTO, focuses on the needs of agencies that are particularly concerned with territorial security and emergencies. Thus GLOSS is improving the KYOTO architecture specifically for what concerns the retrieval, annotation and expert-definition of spatial and geographic data. Most specifically not only concepts but also geographic information can be automatically identified and submitted to experts to be further specified and integrated into GLOSS's knowledge base. The most important new feature is the possibility of associating and retrieving not only point-like geospatial coordinates but also the polygonal extension for geographic entities, something that will in turn allow for more refined map-based queries when assessing the environmental risk of a given region.

In this paper we shall first give a general description of the GLOSS architecture, of the kind of documents it is applied to and of the general and domain document base it relies upon. Subsequently we will describe the linguistic preprocessing of texts and spend some words on the annotation system. Finally we shall describe the two main semantic pipelines of the system: the concept and geo-info pipeline, that extracts candidate concepts and geo-data to be defined by experts thus enriching the domain knowledge bases; and

[1]FP7 ICT-211423 funded under the FP7 http://www.kyoto-project.eu/ - (Vossen et al., 2010)

[2]http://weblab.iit.cnr.it/gloss/node/1

the fact pipeline that uses general and domain knowledge to allow users to access analyzed documents by performing a semantic and geographic search on them.

2. General architecture

The goal of GLOSS is to develop a content enabling system that provides deep semantic search and information access to large quantities of distributed multimedia data that may be relevant to the prevention and management of environmental crises and emergency situations. The system is designed to simultaneously process and search data in different languages and has been currently implemented for Italian and English.
GLOSS data are provided by public and internal documents from the Civic Defence Agencies (Protezione Civile[3]) of the Provinces of Pisa and Livorno, in Tuscany, Italy, that are collaborating with the project as users; English data are currently derived from the FEMA website[4].
Users from the Civic Defence Agencies need to deal with large quantities of information often in a short amount of time. Their source of information lies in internal and public documents reporting previous environmental issues for the relevant areas. These historical reports can be used to predict environmental risk for certain area; moreover in case of emergency in a given area, it can be crucial to be able to efficiently extract information about previous events. Finally, a certain area can be affected by environmental risks of different kind, whose effects in case of emergency are likely to combine.
Available data deal with the prediction, prevention and management of events in the environmental domain and come from reports on environmental risks, natural and man-made calamities, and immediate and long-term assistance plans to help local and state governments as well as individuals. Currently GLOSS has a document base of 480 public documents in English and 211 in Italian.
Typically such data contain descriptions of environmental facts with their geographic and temporal references. Here is an example of the kind of texts to be dealt with:

> Mercoledi 18 Gennaio 2012. Sono 10 i feriti a seguito della violenta esplosione di un metanodotto avvenuta oggi nel comune di Tresana (Massa).
>
> [Wednesday 18th of January 2012. 10 people were wounded in a violent explosion of a methan pipeline today in the commune of Tresana (Massa)]

This text contains a fact that can be described in terms of:

event: explosion

theme: methane pipeline

consequences: 10 wounded people

location: Commune of Tresana - Province of Massa

time: 18/01/2012

[3] www.protezionecivile.it
[4] www.fema.gov

Most crucially the geospatial information does not only imply the reference to geospatial coordinates, but also to areas, such as *Valle dell'Arno* or *Isole dell'arcipelago toscano*. Therefore the system must be able to provide the users with an interface where they can query for facts, for dates and for areas, in order to retrieve events that happened at a certain period and location. This system must be multimodal, that is must represent information both in textual form and by projecting it on detailed maps representing the territory of competence for our partner agencies. Access to the system should be multimodal too, allowing for textual queries as well as map search.
Since it is not to be expected that an automatic fact extraction from texts can do all the job, users must be allowed to integrate and add the information provided with new information.
Following the work done in KYOTO, GLOSS has modeled an architecture that takes this data in input, processes them and annotates them with enriched linguistic, semantic and geospatial information using NLP tools, semantic resources (WordNets) and a geographic database. The processed files are then used to:

1. allow for an enhanced and multimodal search of topics of interest for the community;
2. further enrich the available resources by manually adding new domain concepts to the WordNet and new geographic information to the Geodatabase (e.g. new locations, or new information such as extension to known locations).

Thus after the first process of document acquisition and linguistic analysis the system branches into two separate modules:

1. a fact annotation and retrieval module, that automatically extracts knowledge about concepts and events in space and time that can be queried for by the users through a sophisticated multimodal query system;
2. a knowledge modeling module, organized in the form of a Wiki platform, that helps users to identify and define new concepts and geographic locations to be included in the knowledge base.

The second module was not present in KYOTO and has been created to address the specific needs for geographic knowledge management and mining of the GLOSS users.
The two modules work in parallel, but of course new knowledge that is integrated by the second module will feed the first one in that it is going to be available when new documents are processed or if old documents are reprocessed in the system.
The system rests on general and domain WordNets - lexical semantic databases - in two languages (Italian, English) as well as on a PostGis DataBase[5] that contains information about geographic terms; documents are acquired into a document base that contains documents at all levels of processing.

[5] www.postgis.org

The geographical database, named *GeoBase* in the GLOSS project, is based on a portion of the publicly available *GeoNames*[6] geographical database extended with the possibility to associate a spatial information at any location in the form of a polygon. This extension is based on the geometry type defined by the Open Geospatial Consortium[7] in the OpenGis standard and implemented in the PostGIS extension to PostgreSQL object-relational database.
Both the knowledge modeling and the fact mining module use and feed information into the databases and read and write files into the document base (see Figure 1).

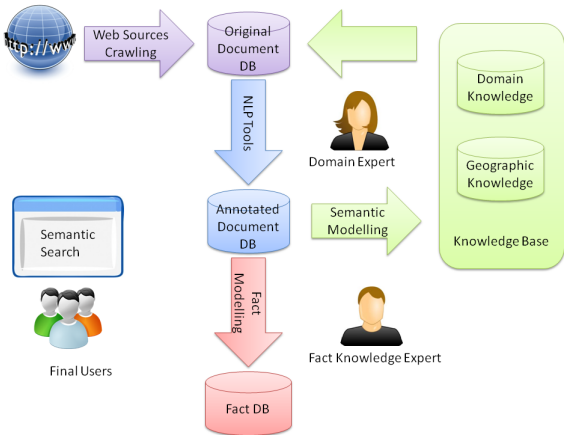


Figure 1: The GLOSS system.

3. Linguistic analysis and annotation format

Documents can be imported into the system in several formats. Currently html pages and PDFs can be imported by a capture module that converts the source into plain text and indexes them in the document base.
Once new documents are entered into the document base linguistic/semantic analysis is automatically performed at the following levels by a pipeline of Natural Language Processing tools relying on statistical and rule based algorithms as well as on the information they retrieve from the knowledge bases in GLOSS:

- morphosyntactic analysis: texts are automatically tokenized, PoS tagged, and syntactically analyzed in order to extract dependencies among constituents
- semantic analysis: semantic classes labeling and word sense disambiguation is performed by accessing information present in the domain and general WordNets
- Geo-Named Entity Recognition and Resolution: references to location occurring in the text are linked to GeoDB entries (at this stage only elements present in the DB are recognized)
- time references recognition: dates are also recognized and formally annotated

[6] www.geonames.org
[7] www.opengeospatial.org

Linguistic pre-processing anchors words and expressions to formal definitions of meaning and uses this information to detect knowledge and facts in text. Standardized annotation of data allows for semantic interoperability of both knowledge and language. For this reason a specific annotation format has been used: KAF (Bosma et al., 2009), an annotation in XML format of linguistic and semantic information. KAF is a layered representation system: every time a new kind of information is added by the processing systems, a new layer is added to the annotated text.
In KAF the following annotation layers are available:

Text: tokenization, sentences, paragraphs, with reference to the source

Terms: portions of Text marked as words and multiwords; they include information on parts-of-speech, declension information, and external references to WordNet senses

Dependencies: relations among Terms

Chunks Terms grouped in constituents and phrases

Geographic information is too complex in GLOSS to be completely inputted into each document, so an external reference is used that points to the *GeoBase*.

4. The GLOSS Knowledge Module

Knowledge modeling is performed into two steps: automatic extraction of candidate elements and knowledge integration by domain users. The first step is composed by two algorithms:

- term extraction and semantic annotation: simple and complex terms that are linked to their relevant synsets in the WordNets; candidate terms are also extracted using special heuristics (by TEA - term extracting agents) and proposed to the users for further integration in the knowledge base, so that they can be later recognized in the analysis of new texts.
- geographic terms recognition: known geographic locations are already linked to their entries in a geospatial database that has been extended from GeoNames in the pre-processing phase; furthermore semantic analysis carries information about generic geographic terms in the text (such as *valley*, *river*, ...); by combining these two levels of information the geographic extraction agent (GEA) can extract candidate location terms (e.g *valley of the river Serchio*) that identify areas of potential interest. Similarly to what happens for the terms, candidates are submitted to users to be evaluated and added to the database with additional information (such as the area as a polygonal shape).

The TEA and GEA algorithms rely on rules that are applied to the chunked and dependency parsed annotated documents; they first retrieve nominal heads, then elements that modify these heads (adjectives or prepositional phrases) and modifiers of modifiers (if any). By doing this they

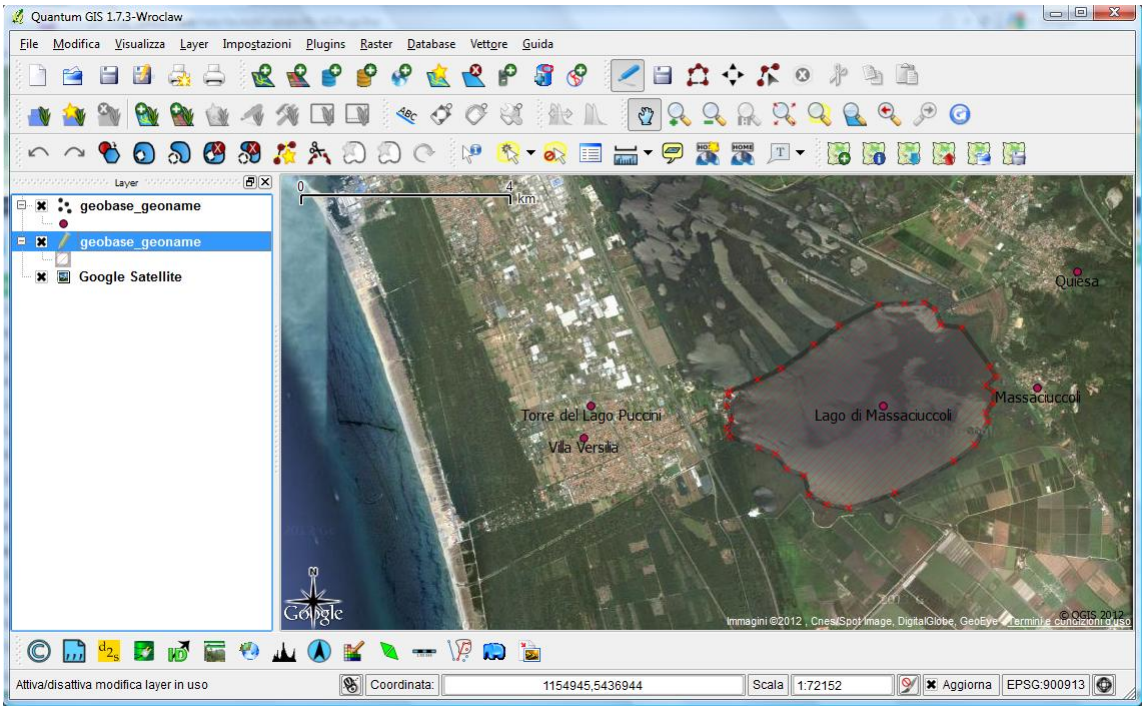


Figure 2: The GeoBase editing tool developed in GLOSS. It allows for advanced editing, by seamlessly interfacing to a GIS (such as Qgis).

can retrieve chains of terms with different levels of specificity, such as *Insurance* > *Flood Insurance* > *Flood Insurance for Mortgage Lenders*. In order to help the knowledge user, the frequency of occurrence for each individual and complex term in the documents is recorded, and *log likelihood* is calculated to measure the association strength within complex terms. Also, the algorithm checks for correspondences with the generic and domain WordNet.

Knowledge integration is performed by users with the help of a Wiki editor, along the line of the model proposed in Kyoto. Candidate terms and locations are proposed by the system to the expert user that can graphically access the existing knowledge in an editing environment (Abrate and Bacciu, forthcoming). In addition to what was proposed in KYOTO, not only does the user need to integrate terms as concepts into the WordNets, but also add new locations and draw areas for known ones via a graphic tool (see Figure 2).

5. The GLOSS fact pipeline and semantic search

The fact miner uses information from the linguistic analysis to extract semantic relations between concepts. For instance agentive relation is extracted from the noun verb dependencies. Most relations are directional, for instance person $->$ AGENT $->$ breathe. In our example in Figure 3 the main relations are: when, where, agent, qualification and modification.

Facts are annotated on the texts as an additional layer of information and can be used to retrieve information. *Semantic search* is performed by an interface allowing users to refine queries to the document base by accessing both the concept DB (selecting senses on WordNets) and the Geo DB (selecting an area).

Users can start the search by inserting some words in the search-box provided. The system sends back, along with the results, its interpretation of the query, clarifying which WordNet sense has been used to retrieve the documents.

For example, if the query contains the word “flood”, the system could reply:

> Results assuming “flood” is:
>
> [1]: (n) flood, inundation, deluge, alluvion (the rising of a body of water and its overflowing onto normally dry land) “plains fertilized by annual inundations”
>
> `- list of results for [1] -`
>
> “flood” may also refer to:
>
> [2]: (n) flood, inundation, deluge, torrent (an overwhelming number or amount) “a flood of requests”; “a torrent of abuse”
>
> [3]: (n) flood, floodlight, flood lamp, photoflood (light that is a source of artificial illumination having a broad beam; used in photography)
>
> [4]: (n) flood, overflow, outpouring (a large flow)
>
> [5]: (n) flood, flowage (the act of flooding; filling to overflowing)
>
> [5]: (n) flood tide, flood, rising tide (the occurrence of incoming water (between a low tide and the following high tide))

If users want to change the interpretation they can choose one of the listed alternatives.

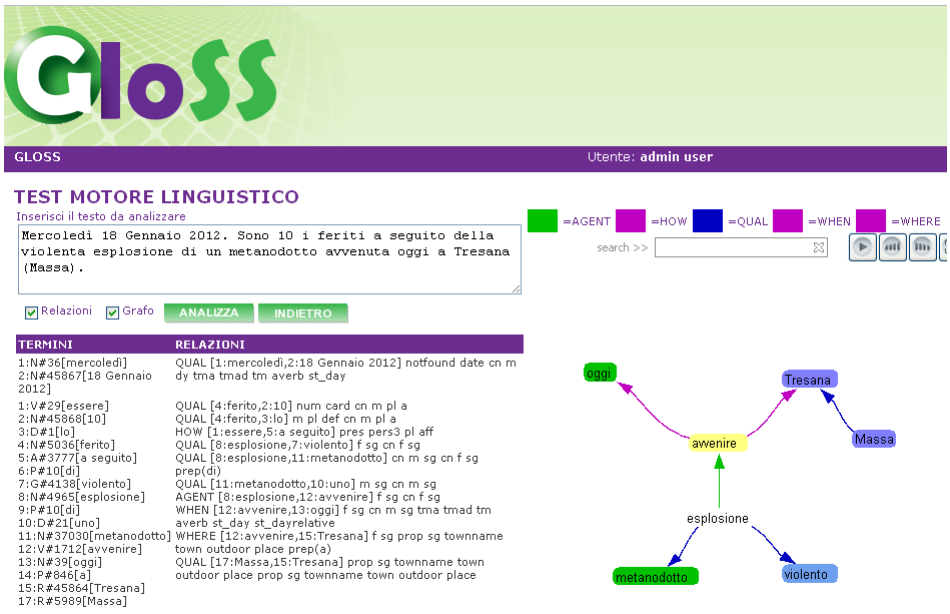


Figure 3: GLOSS fact extractor.

Let it be stressed here that the Knowledge Base that is available to the Semantic Search module has been previously enriched as described in the Knowledge Module; future work will verify whether specific knowledge present in the domain WordNet allows for better disambiguation and extraction results.

The results of the search are shown in a list containing excerpts of the documents from which they have been extracted. The user can click on one excerpt and read the whole corresponding document.

The query tries to collect all factual relations for to the searched terms, in this case for instance all facts concerning floods, with a special attention to the location. This allows for the results to be also displayed in a map as placemarks showing their localization. If the geographic search is enabled, the results can be filtered by retrieving only the documents whose localization lies in the current map viewport. Navigating the map by panning and zooming is also a way to further filter the results.

6. Conclusion

We described GLOSS, an architecture for the analysis and semantic mining of document in the public safety domain. This system, combining state of the art natural language and semantic processing, a refined geographic data management system and methods for semi-automatic collaborative definition of knowledge by experts, can be used by organizations in the public safety domain to efficiently organize, increment and access their domain knowledge, which in turn can be crucial to collect the necessary information for decision taking in emergency situations.

7. Acknowledgements

The work described here is funded by GLOSS, a project within the Regione Toscana Bando Unico R&S anno 2008 - linea A.

8. References

M. Abrate and C. Bacciu. forthcoming. Visualizing word senses in WordNet Atlas . In *In Proceedings of LREC 2012*, Istanbul, May 21-27 , 2012.

C. Aliprandi, F. Ronzano, A. Marchetti, M. Tesconi, and S. Minutoli. 2011. Extracting events from wikipedia as rdf triples linked to widespread semantic web datasets. In *Proceedings of the 4th international conference on Online communities and social computing*, OCSC'11, pages 90–99, Berlin, Heidelberg. Springer-Verlag.

W. Bosma, P. Vossen, A. Soroa, G. Rigau, M. Tesconi, A. Marchetti, M. Monachini, and C. Aliprandi. 2009. Kaf: a generic semantic annotation format. In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon (GL 2009)*, pages 145–152, Pisa, Italy, September 17-19, 2009.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Open Geospatial Consortium Inc. 2011. Opengis implementation standard for geographic information - simple feature access v1.2.1.

P. Vossen, W. Bosma, E. Agirre, G. Rigau, and A. Soroa. 2010. A full knowledge cycle for semantic interoperability. In A. Fang, N. Ide, and J. Webster, editors, *Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation, in conjunction with the Second International Conference on Global Interoperability for Language Resources (ICGL 10)*, Hong Kong, January 15-17, 2010.

Requirement Extraction from Transport Policy Documents

Oliver Mason, Aleksandar Trklja, Dominik Vajn
University of Birmingham
United Kingdom
O.Mason@bham.ac.uk

Abstract

Requirements are an important concept in systems engineering. We present an approach to the automatic extraction of requirement statements from transport policy documents using a local grammar. The grammar has been developed using standard corpus linguistic methods. With a fairly straight forward grammar we can identify instances of requirements, as they are expressed using a small number of distinct surface patterns.

One additional complication involved in this research is the issue of clearly separating requirements from policies or strategies, which are generally at a higher level. We have identified a number of different patterns that can help in distinguishing between such statements.

Keywords: local grammar, policy, requirements

Introduction

Policy documents are a means for political decision makers to communicate with those who implement the policies, such as other government agencies or private sector companies bidding for contracts. They define a framework which guides the implementation of what is required, without actually prescribing it in detail: the details of the technical realisation are left to the experts who decide how to best implement the specified requirements.

However, policy documents at present are far from ideally suited for this purpose. Instead, they will generally contain a mixture of policies, strategies, requirements, and even specifications (which arguably should not be contained in such high-level documents). Requirements, which are what implementers are interested in, need to be extracted from the documents by reading through and deciding which statements are relevant for the purpose of a target audience of systems engineers.

This raises several issues: Not only is it a rather time-consuming process to read through long documents to pick out a few relevant sentences, it is also not guaranteed that any two engineers would identify the exact same requirements from reading the same document.

The current project, a collaboration with the Birmingham Centre for Railway Research and Education, attempts to resolve those questions by designing a computer system capable of extracting requirement statements from (transport) policy documents. This is done through a description of the surface grammatical structures in which such statements are phrased using a formalism called 'local grammar' (Gross 1993, Sinclair and Hunston 2000).

Describing the phraseology of requirement statements does, though, rely on an existing description of such structures, so a further component of the current research is to identify how exactly requirements are expressed in those policy documents. Given the lack of such an *a priori* description, automatic approaches such as machine learning cannot be applied.

To summarise, the aim of the project is to manually identify linguistic patterns used to realise requirement statements in policy documents, and then express those patterns in a form that can be automatically identified by a pattern recogniser. The focus with the identification process is on recall rather than precision, as it is tantamount to identify all *potential* requirements: it is easily possible for an end user to reject a statement as a false positive, whereas any requirements that have not been flagged up will be lost in the text.

Requirements in Systems Engineering

Systems Engineering (SE) is a structured process to the realisation of (complex) systems. A number of models

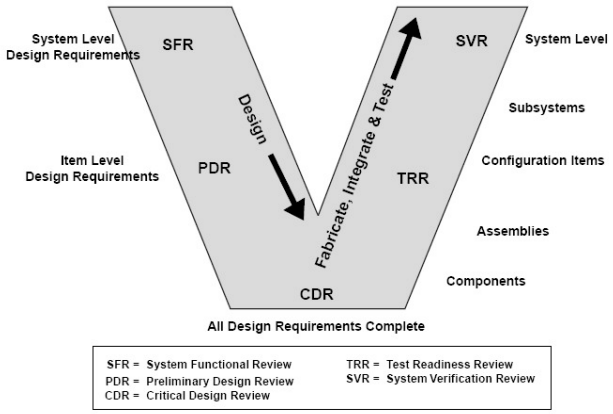


Fig 1: The V-model in Systems Engineering (see N.N. 2001)

have been developed in SE, such as the V-model (see fig. 1), which describes the development of a system chronologically (from left to right) and by level of detail (top – abstract; bottom – specific). Policy documents would be located in the top-left corner, as they stand at the beginning of the system creation and describe the system on a very abstract level, (ideally) devoid of any references to implementation-specific details.

Requirements extend a bit further along the V-shape, as there are different levels of specificity. Terminology in SE is rather unclear and different authors use the same terms to describe different concepts; in principle requirements have to be distinguished from strategies and policies.

Despite all attempts in SE to create systematic approaches to identifying requirements, this is still 'something of a black art' (Chris Bouch, personal communication). By specifying a set of linguistic structures that typically express requirements, we will attempt to support research into making their identification more objective and reliable.

Transport Policy Documents

The policy documents we are investigating are freely available as PDF files; these can then be converted into plain text using a range of conversion tools. The main issue with the conversion is, however, that those documents typically have a less straight forward layout in that there will be page headers and titles which interfere with the plain text. Initially we have decided not to do any manual post-editing, but we will have to revisit this point once we have evaluated how much this hinders the automatic analysis, as the text is occasionally interrupted in mid-sentence when there is a page break. At the present stage this is not a priority, as we can initially assume that we are working with plain text only.

The texts were split into sentences using a rule-based sentence splitter, and for pattern matching they were tokenised, lemmatised and POS-tagged (using QTag, Mason 2003).

Policy documents are produced at several political levels. The initial data set comprises documents issued by the UK Department of Transport, but there are similar documents available from the EU (such as the white paper on transport policy). While analysing the data, a portion of the data set is excluded to be used for evaluation at a later stage. While there are a few documents only, they are generally between 20K and 50K words in length.

Linguistic Structures

We followed several parallel strands of analysis with the overall aim to identify how requirements are expressed. To guide the analysis, a pre-annotated document was available where a systems engineer had marked up what he thought were requirements. The distinction between requirements on the one hand and policies/strategies on the other was made on the basis that strategies express more broad and long term goals, whereas requirements are more specific in their goals and also their addressees.

The first step was to come up with a basic description of how the linguistic concept of 'requirement', which would then be used for an automatic analysis of a further document. A close reading of the processed document was then undertaken to evaluate the coverage of the initial description and to identify any further patterns that had been missed.

Three approaches were combined for the initial analysis: keywords analysis, parts-of-speech-based filtering, and synonyms analysis.

Keyword analysis is conceptually identical to collocation; instead of using the local environment of a node

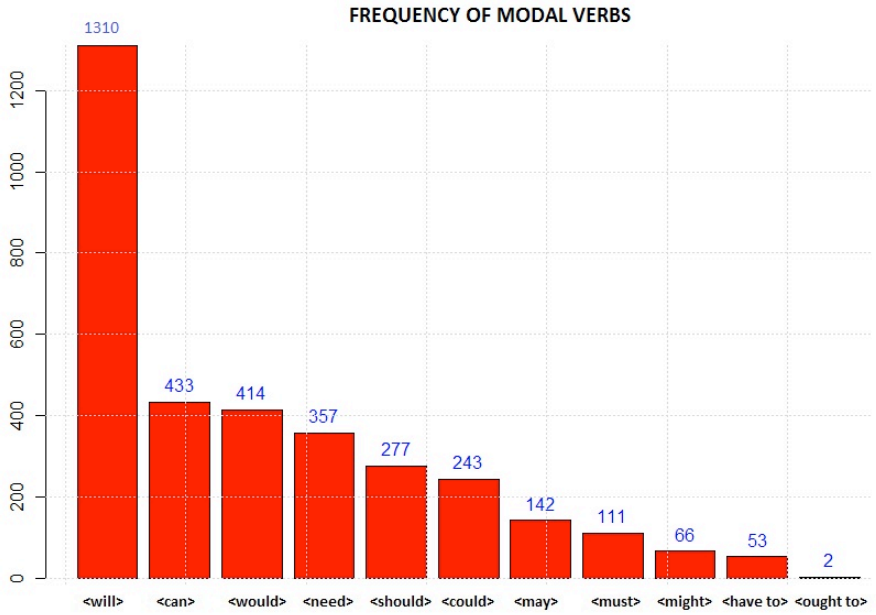


Fig 2: The distribution of modal verbs across policy documents

word for evaluating the salience of the collocates, a whole text is used as the sample which is compared with the reference corpus. This provides a list of words which occur in the particular text with a frequency higher than expected from their distribution in a general corpus.

Keyword analysis is implemented in WordSmith tools (Scott 2008); the British National Corpus was used as a reference corpus. This method proved to be of very limited use as the highest 'keyness' values were found with words (mostly nouns) that refer to various transport-related topics (*airport*, *railway*, *transport*, *passenger*, *train*, *Heathrow*, *infrastructure*). The keyword analysis is concerned with the 'aboutness' of texts (Scott & Tribble 2006) and since our texts are about different forms of transport, airports, passengers,roads, and traffic it is not very surprising that these words are most dominant.

Interestingly, requirements are sometimes referred to as 'shall-statements' by systems engineers, though the policy documents we have so far consulted do not contain a single instance of *shall*. Figure 2 shows the distribution of modal verbs in all documents.

For part-of-speech-based filtering the corpus is first tagged using QTag (Mason 2003) and then indexed with the IMS Corpus Workbench (Evert 2005). The CQP query language made it possible to determine the frequency of all lexical items in relation to grammatical categories. Verbs present the most interesting category because here we find at the top of the list such lexical items as *need*, *require*, *want*, *seek*, *expect* and *encourage* that seem to be relevant for the study of requirements. By looking at the lexico-grammatical profile of the identified verbs it was possible to identify first patterns that express requirements (e.g. **NP *need to* VERB** or **NP *be required to* VERB**). In addition, we exploited the distributional hypothesis which claims that lexical items with similar meaning will occur in similar textual environments (Harris 1970) and studied the collocates that occur with the previously identified verbs. Thus, the verbs *ensure*, *improve*, *reduce*, *provide* and *deliver* occur frequently with the constructions *need to* and *be required to* in our corpus and we expected to find some other semantically similar expressions. This method was successful to some extent and for example we found that ***maintain* NP** which occurs with ***be required to*** also occurs with ***it be essential*** and **NP *seek to***. These two constructions proved to be very productive in our corpus because they can be combined with other lexical items to express requirements.

Finally, we also looked at the thesauri or thesaurus-like lexical databases such as WordNet (Fellbaum 1998), Collins Thesaurus (2002) and Longman Language Activator (2002) in order to find words that express simi-

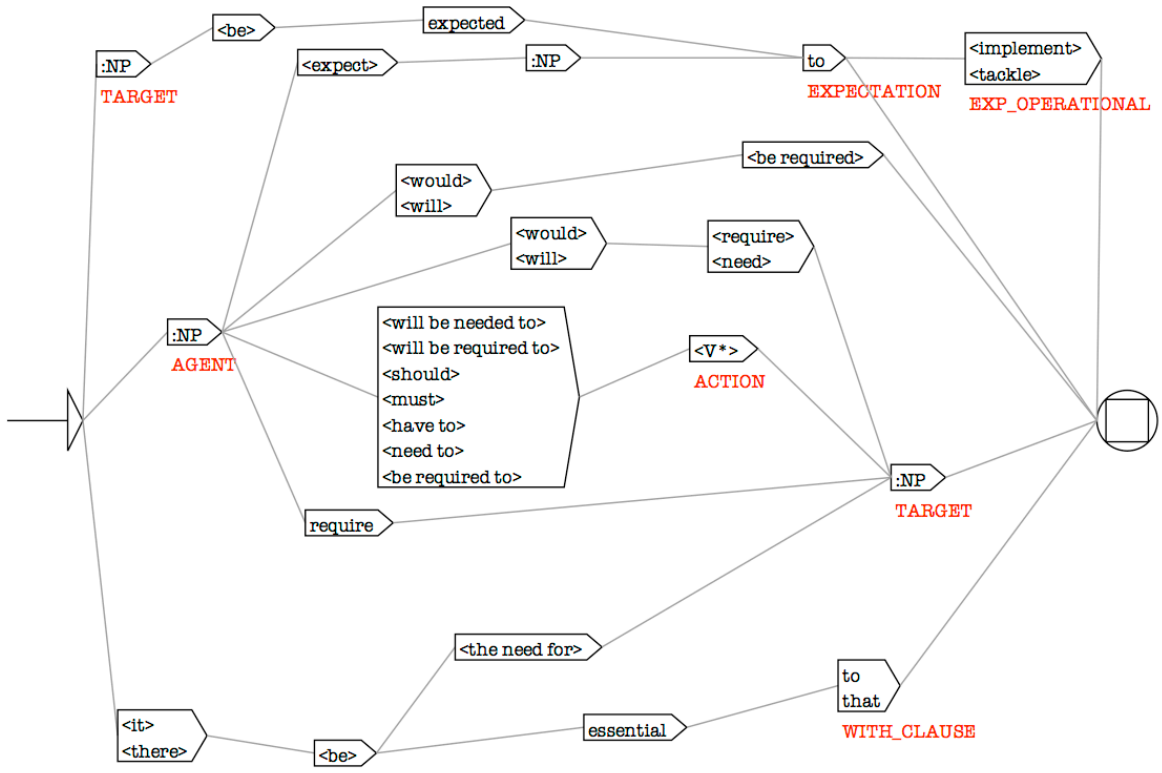


Fig 3: the initial local grammar of requirements

lar meaning. This method confirmed some of the previously identified lexical items and pointed out some new candidates.

To conclude, using part-of-speech-based filtering and synonym analysis have proven more effective for our purposes than keyword analysis.

Analysis

From the initial analysis of the linguistic structures we have derived the following local grammar (see fig. 3)

The local grammar approach is very flexible and it makes possible to put semantic labels on the lexical items associated with a pattern that expresses specific meaning. Thus, one important pattern in our data contains the lexical items <be required to>, <need to>, <be needed to>, <should>, <must> and <have to>. All these lexical items are preceded by a noun phrase and followed by a verb and if the verb is transitive by another noun phrase. Because they express similar meaning and occur in the same pattern we can say that they create one semantic set that can be labelled as {NEED TO}. In addition, the noun phrases can be labelled as {AGENT}, verbs as {ACTION} and the second noun phrase as {TARGET}. An analysis has been conducted to find out if the lexical items that belong to one of the latter groups can be further specified. So far we were able to distinguish between two types of constructions that belong to {AGENT}: {TRANSPORT_TEAM} and {AUTHORITIES}. The former is realized through NPs such as <airport operators> and the latter as <local government> or <the Northern Ireland authorities>. This information can be useful because the categories help in the process of separating requirements from policy or strategy statements.

Having analysed the set of policy documents with the local grammar we then went through the complete documents to evaluate the level of coverage. A few further patterns were identified that had not been matched by the local grammar, but could easily be integrated. This quickly pushed up the recall rate and in the remainder of the data no new structures could be found.

The Language of Requirements

According to our findings, the language of requirements is associated with a limited number of lexical items that occur in specific patterns. In this section we will list a number of relevant patterns with example from the corpus. All of these patterns can easily be recognised using a local grammar.

'require'

(agent) <be> *required* to-inf

- Transport will *be* ***required*** *to contribute* towards achieving this ambitious...
- Operators will *be* ***required*** *to participate* in it and to provide the necessary...
- Network Rail will *be* ***required*** *to deliver* infrastructure capacity for a specified...

(goal) *requires* (condition)

- A well-performing transport network ***requires*** substantial resources...
- …so effective action ***requires*** strong international cooperation...
- …the potential of private finances equally ***requires*** an improved regulatory framework...
- The completion of the TEN-T network ***requires*** about €550 billion until 2020...

***will* + *require* + (goal)**

- To meet this goal will also **require** appropriate infrastructure to be developed...
- It will also **require** an investment plan for new navigation...
- The need to increase capacity in some areas *will* ***require*** us to consider a range of solutions...

modality

(goal) *must* + *be* V-ed + (specification, agent)

- The environmental record of shipping can and ***must*** *be improved* by both technology and better…
- Better rail airport connections ***must*** *be devised* for long distance travel…
- Transport charges and taxes ***must*** *be restructured* in the direction of wider…
- Project assessment and authorisation ***must*** *be carried out* in an efficient and transparent…

(agent) *must* + V + (goal)

- The core network ***must*** *ensure* efficient multi-modal links between...
- Future development ***must*** *rely* on a number of strands:...
- eligible for EU funding ***must*** *reflect* this vision and put greater emphasis on...

(goal/aim) *should* + *be* + V-ed

- Urban Mobility Plans ***should*** *be fully aligned* with Integrated Urban Development...
- Fuel saving techniques ***should*** *also be developed* and promoted in other modes...
- information technology tools ***should*** *be widely deployed* to simplify administrative…

(aim) *should* + *be* + to-inf

- Our aim ***should*** *be to become* the safest region for aviation...

(agent/object) should + V

- The plans ***should*** *address* the issue of prioritisation...
- (high speed) rail ***should*** *absorb* much medium distance traffic...
- The EU aviation industry ***should*** *become* a frontrunner in the use of low-carbon fuel...

Predicative Structures

(goal) + <be> + adj

- Their development ***is*** *vital* to handle increased volumes of freight both…
- Mobility ***is*** *vital* for the internal market and for the quality of life…
- It ***is*** *therefore important*, besides encouraging alternative transport solutions…
- Setting the framework for safe transport ***is*** *essential* for the European citizen supervision of safety certification are essential in a Single European Railway Area.
- Innovation ***is*** *essential* for this strategy
- Economic models and rail forecasting tools ***are*** *essential* to project appraisal and can assist forward…
- It ***will be*** *important* to align the competitiveness and the social agenda…
- Our international networks ***are*** *also vitally important*…

(goal) + <be> + *key*

- vehicles and traffic management ***will be*** *key* to lower transport emissions

Strategy Patterns

The following are examples of patterns that do *not* express requirements, but are used with more high-level strategies:

(agent) *want* + to-inf

- We ***want*** *to encourage* low-carbon technology...
- we ***want*** *to encourage* the use of the railways as a lower-carbon..
- We still ***want*** *to cut* transport's carbon footprint...

(agent) + *seek* (NP/to-inf)

- ...we will ***seek*** *solutions* that mitigate unavoidable adverse...
- We will ***seek*** *to mainstream* this approach as we develop our...

(agent) + *expect* + to-inf

- ...we would ***expect*** *to identify* a reduced need for major new transport...
- We ***expect*** *to be able to* make progress against all five...
- we ***expect*** *there to be* a strong synergy between different goals...

***aim* + <be> + to-inf**

- Our *aim* ***is*** *to ensure* that we have a transport system that...

Evaluation

Evaluating the results of the analysis is problematic due to the circularity involved: we define what we mean by 'requirement', formalise the definition in a local grammar, and then recognise them in the data. Given the issues mentioned above about subjectivity and the 'black art' of identifying requirements in a policy document it is necessary to use a group of systems engineers to assess the accuracy of the recognised statements. We have not yet been able to do this, but such an evaluation exercise has been planned and will be reported in subsequent publications.

Unfortunately the evaluation is thus rather impressionistic. We have been able to describe a local grammar expressing requirements in a straight forward way by using standard methods of corpus analysis to identify relevant surface structures. Our findings are consistent with some of the basic tenets of corpus linguistics, in that there is a correlation between form and meaning, i.e. certain surface structures correspond to certain meanings. There is a fairly limited set of expressions that can be used to cover all those statements that a linguist would classify as 'requirement'. It remains to be seen if systems engineers would agree with that classification.

Summary

We have shown the process by which a local grammar of requirements has been developed, though the work is not yet completed as a full evaluation has not taken place so far. We are, however, confident that the automatic recognition process works reliably for the purpose for which it was designed, namely to support research in systems engineering on how requirements are expressed in the English language, and to build a system that enables engineers to access the requirements without having to consult the actual policy document in full.

The main aim of the project was to formalise the way requirements are expressed in the language of policy documents; while we have worked on transport policy documents only, the principle should be transferable to any other domain, as the local grammar does not include any domain specific vocabulary. Part of the planned future work is to apply the system to a broader range of relevant documents.

A system that can extract requirements from policy documents in a fully automated fashion has a number of possible applications: not only can it support engineers needing to design systems (as they would not need to read the full policy document but would simply have a list of requirements ready-made), but it can also aid the writers of policy documents by giving them an indication of what requirements are contained in a document (and thus allows checking the document for completeness/correctness).

A further plan is to apply the system to policy documents in languages other than English by using machine translation as a first preprocessing step. If the text is also available in English (as would be the case with EU policy documents) the output of this process could then be directly evaluated to see how well such documents fare with automatic translation.

References

Collins English Dictionary & Thesaurus (2002). HarperCollins Publishers Ltd., Version 3.0 Software

Evert, S. (2005). The CQP Query Language Tutorial. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/cqp-tutorial.pdf

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Bradford Books.

Gross, M. (1993). Local Grammars and their Representation by Finite Automata. In M. Hoey (Ed.), Data, Description, Discourse, London: HarperCollins, pp. 26-38.

Harris, Z. (1970). Distributional structure. In Papers in structural and trans-formational Linguistics, pp. 775-794.

Longman Language Activator (2002). Harlow: Longman.

Mason, O. (2003). QTag. http://phrasys.net/uob/om/software.

N.N. (2001). Systems Engineering Fundamentals. Defense Acquisition University Press

Sinclair, J. & Hunston, S. (2000). A Local Grammar of Evaluation. In S. Hunston, G. Thompson, G (Eds.) Evaluation in Text: Authorial Stance and the Construction of Discourse, Oxford: OUP, pp. 74-101.

Scott, M. (2008). WordSmith Tools version 5, Liverpool: Lexical Analysis Software.

Scott, M. & Tribble, C. (2006). Textual Patterns: Key words and corpus analysis in language education. Amsterdam: John Benjamins.