CORPORA DOCUMENTATION
Local Government Documentation

1 BASIC INFORMATION

1.1 Corpus composition

This corpus is a collection of different governmental resources, containing two types of documents: minutes, which were taken during local council meetings (covering the years from 2007 till 2010) and memorandums (covering from 2008 till 2011).

This corpus, consisting of raw text files and comma separated values (CSV) files, is the percentage that could be extracted from the original corpus.

Some issues arise due to Maltese characters. It is important to note that not all documents contain the right Maltese characters. Some documents may replace:

ġ -> g ; ż -> z ; ħ -> h ; ċ -> c

With those being on the right hand side also Maltese characters, except *c*. Furthermore, in some of the documents, the keyboard equivalence of the character is printed, rather than the character itself (and this is also dependent on whether the user made use of the 47 or 48-key keyboard layout).

1.2 Representation of the corpora (flat files, database, markup)

The corpus is organized in two folders:

a. TXT – this folder contains a collection of text files; minutes and memos which were extracted from the original corpus.

b. CSV – this folder contains a collection of CSV files, most of which contain financial information, extracted from excel files in the original corpus

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: Dr. Alexiei Dingli
Affiliation: University of Malta, Department of Communications & Computer
Engineering
Position: Senior Lecturer, Department of Intelligent Computer Systems
Telephone: +356 2340 2486
e-mail: alexiei.dingli@um.edu.mt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform

2.3 Copyright statement and information on IPR

META-SHARE Commons BY NC SA

3 TECHNICAL INFORMATION

3.1 Directories and files

2 Directories:

a. TXT directory – 4402 text files

b. CSV directory – 1275 CSV files

3.2 Data structure of an entry

n.a.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)

TXT directory size on disc: 53.5MB

CSV directory size on disc: 184 MB

Corpus size on disc: 238MB

4 CONTENT INFORMATION

4.1 Type of the corpus (monolingual/multilingual, parallel/comparable, raw/annotated)

The data is mostly monolingual but is in two languages - both English and Maltese. The corpus is raw.

4.2 The natural language(s) of the corpus

Maltese and English

4. 3 Domain(s)/register(s) of the corpus

Local councils' meeting minutes, government memorandums and financial information (with respect to decisions undertaken by the local councils).

4.4 Annotations in the corpus (if an annotated corpus)

4.4.1 Types of annotations (paragraph mark-up, sentence mark-up, lexical mark-up, syntactic mark-up, semantic mark-up, discourse mark-up)

n.a.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),

n.a.

4.4.3Alignment information (if the corpus contain aligned documents: level of alignment, how it was achieved)

n.a.

4.4.4 Attributes and their values (if annotated)

n.a.

4.5 Intended application of the corpus

Information extraction techniques 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.

4.6 Reliability of the annotations (automatically/manually assigned) – if any

n.a.

5 RELEVANT REFERENCES AND OTHER INFORMATION