MLRS Corpus

1 BASIC INFORMATION

1.1 Corpus composition

142,397 corpus texts from 10 genres

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

txt files with XML-like tags for texts, paragraphs, sentences, POS tags

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

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

Albert Gatt
Institute of Linguistics
University of Malta
albert.gatt@um.edu.mt

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

zip. file for download in the META-SHARE platform

2.3 Copyright statement and information on IPR

META-SHARE NonCommercial NoRedistribution signed by Albert Gatt

3 TECHNICAL INFORMATION

3.1 Directories and files

The file "corpus.zip" expands into a folder "corpus", containing the file "tagged.zip", which expands into the folder "cwb.final". This folder contains the files:

- filelist.txt
- malti02.academic.txt
- malti02.law.txt
- malti02.literature.txt
- malti02.metadata.txt
- malti02.misc.txt
- malti02.parl.txt
- malti02.parl.txt.bak
- malti02.press.txt
- malti02.religion.txt
- malti02.speeches.txt
- malti02.web.genral.txt
- malti02.web.wiki.txt
- README.txt
- removed-from-corpus.txt
- tend.txt
- tstart.txt

3.2 Data structure of an entry

Each text in a text file is marked by <t>…</t>, each paragraph … , each sentence <s>…</s>, each word as a single entry per line, followed by a tab and its tag, giving the structure:

```
<text id="kts10">
<p id="0">
<s id="0">
Dahla    NN
</s>
</p>
<p id="1">
<s id="1">
Il-      DDC
Gawgaw           VV
```

u CC
l- DDC
Imlejka NN
</s>

…
</text>

3.3 Corpora size (nmb. of tokens, MB occupied on disk)
10 subcorpora
124,727,981 tokens
725,487 types

4 CONTENT INFORMATION

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

4.2 The natural language(s) of the corpus
Maltese (some code switching with English and Italian can occur in literature and parliamentary debates)

4. 3 Domain(s)/register(s) of the corpus
parliament, press, academic, misc, web, law, literature, speeches

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)
All texts of a genre are in one .txt file for that genre. In this file, texts are marked <t>…</t>, paragraphs are marked … , sentences are marked <s>…</s>, and one word per line, followed by a tab and its POS tag.

4.4.2 Tags (if POS/WSD/TIME/discourse/etc –tagged or parsed),
tagset can be viewed here: http://mlrs.research.um.edu.mt/index.php?page=34

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)
see 4.4.1 and 4.4.2

4.5 Intended application of the corpus
To be used in the corpus workbench of the CQPweb system (http://cwb.sourceforge.net/doc_links.php)

4.6 Reliability of the annotations (automatically/manually assigned) – if any
Automatically tagged using the TnT tagger (http://www.coli.uni-saarland.de/~thorsten/tnt/), trained on manually annotated texts (ca. 26,000 tokens), before being applied to the whole corpus. Accuracy is at around 96%.

5 RELEVANT REFERENCES AND OTHER INFORMATION
http://mlrs.research.um.edu.mt/index.php?page=34