Maltese Acquis Communautaire MT

1 BASIC INFORMATION

1.1 *Corpus composition*

This is the Maltese version of the *Acquis Communautaire* (AC), which is the total body of European Union (EU) law applicable in the EU Member States. It consists of selected texts between the 1950s and today.

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

The corpus consists of folders (named by year) and xml files (containing the law texts).

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Ralf Steinberger
European Commission - Joint Research Centre (JRC)
Address: Via Fermi 2749, 21027 Ispra (VA), Italy
URL: http://langtech.jrc.ec.europa.eu/
Telephone: +39 0332 78-5648 or 78-9478
Email: Ralf.Steinberger@jrc.ec.europa.eu
Fax: +39 0332 78-5154

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

The resource will be uploaded on the META-SHARE platform.

2.3 Copyright statement and information on IPR

The corpus is freely available from the JRC website. The legislative texts contained in it are in the public domain. The use of the corpus demands the attribution to the European Communities and to the source. Also the dislaimer “Only European Community legislation printed in the paper edition of the Official Journal of the European Union is deemed authentic.” is to be retained in the xml files of the corpus.
When translations are made from a law text in the corpus, they have to be accompanied by another disclaimer:
“For the reasons stated in the disclaimer above, it is advisable to ensure that translations are made from the printed, authentic version of the Official Journal. This precaution, while minimizing the risk of error, does not confer any legal status whatsoever to the translated text. The following notice shall accompany the translated text, printed below the acknowledgement: 'Originally published in the official languages of the European Union in the Official Journal of the European Union by the Office for Official Publications of the European Communities. Responsibility for the translation into [specify language] from the original [specify language] edition lies entirely with [name of translation copyright holder].' Moreover, please note that we do not consider a "further commercial dissemination" the inclusion, as reference material for consultation purposes, of small amounts of relevant legislative texts in articles/thesis/studies/reports/books issued by third-party authors or publishers, whatever the means, and disseminated subject to payment. “

(see http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html#Usage)

3 TECHNICAL INFORMATION

3.1 *Directories and files*

48 directories (each directory for a year) containing 10550 files

3.2 Data structure of an entry

Not relevant as the corpus consists of text files (XML-structured) with the lowest tagged entry in each file being a paragraph.

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

20,926,909 tokens, taking 314 MB on disk

4 CONTENT INFORMATION

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

monolingual, raw, annotated for paragraphs

4.2 The natural language(s) of the corpus

Maltese

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

Legalese

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)

The files contain annotations at paragraph level.

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

For paragraphs, … tags are used.

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

There are no alignment files in the monolingual corpus files. Alignment was done on paragraph level for 231 language pair combinations by the JRC, as pointed out here: http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html#Alignment

4.4.4 Attributes and their values (if annotated)

Only … tags for paragraphs.

4.5 Intended application of the corpus

Machine translation by aligning two language versions (e.g. Maltese and English).

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

unknown

5 RELEVANT REFERENCES AND OTHER INFORMATION
http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/licence.html

Maltese Acquis Communautaire EN

1 BASIC INFORMATION

1.1 *Corpus composition*
This is the English version of the *Acquis Communautaire* (AC), which is the total body of European Union (EU) law applicable in the EU Member States. It consists of selected texts between the 1950s and today.

1.2 Representation of the corpora (flat files, database, markup)
The corpus consists of folders (named by year) and xml files (containing the law texts).

1.3 Character encoding
UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Ralf Steinberger
Name: European Commission - Joint Research Centre (JRC)
Address: Via Fermi 2749, 21027 Ispra (VA), Italy
URL: http://langtech.jrc.ec.europa.eu/
Telephone: +39 0332 78-5648 or 78-9478
Fax: +39 0332 78-5154

2.2 Delivery medium (if relevant; description of the content of each piece of medium)
The resource will be uploaded on the META-SHARE platform.

2.3 Copyright statement and information on IPR
The corpus is freely available from the JRC website. The legislative texts contained in it are in the public domain. The use of the corpus demands the attribution to the European Communities and to the source. Also the dislaimer "Only European Community legislation printed in the paper edition of the Official Journal of the European Union is deemed authentic." is to be retained in the xml files of the corpus.
When translations are made from a law text in the corpus, they have to be accompanied by another disclaimer:

“For the reasons stated in the disclaimer above, it is advisable to ensure that translations are made from the printed, authentic version of the Official Journal. This precaution, while minimizing the risk of error, does not confer any legal status whatsoever to the translated text. The following notice shall accompany the translated text, printed below the acknowledgement: 'Originally published in the official languages of the European Union in the Official Journal of the European Union by the Office for Official Publications of the European Communities. Responsibility for the translation into [specify language] from the original [specify language] edition lies entirely with [name of translation copyright holder].' Moreover, please note that we do not consider a "further commercial dissemination" the inclusion, as reference material for consultation purposes, of small amounts of relevant legislative texts in articles/thesis/studies/reports/books issued by third-party authors or publishers, whatever the means, and disseminated subject to payment. “

(see http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html#Usage)

3 TECHNICAL INFORMATION

3.1 *Directories and files*
47 directories (each directory for a year) containing 23,551 files

3.2 Data structure of an entry
Not relevant as the corpus consists of text files (XML-structured) with the lowest tagged entry in each file being a paragraph.

3.3 Corpora size (nmb. of tokens, MB occupied on disk)
34,588,383 tokens, taking 468 MB on disk

4 CONTENT INFORMATION

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

4.2 The natural language(s) of the corpus
English

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

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)

The files contain annotations at paragraph level.

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

For paragraphs, … tags are used.

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

There are no alignment files in the monolingual corpus files. Alignment was done on paragraph level for 231 language pair combinations by the JRC, as pointed out here: http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/README_Acquis-Communautaire-corpus_JRC.html#Alignment

4.4.4 Attributes and their values (if annotated)

Only … tags for paragraphs.

4.5 Intended application of the corpus

Machine translation.

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

unknown

5 RELEVANT REFERENCES AND OTHER INFORMATION

http://optima.jrc.it/Acquis/JRC-Acquis.2.2/doc/licence.html