Illum Corpus

1 BASIC INFORMATION

1.1 *Corpus composition*
The full editions of ILLUM from 12/11/2006 to 30/05/2010 (185 issues).

1.2 Representation of the corpora (flat files, database, markup)
XML files with paragraph marking (<paragraph> … </paragraph>) and each word on a separate line.

1.3 Character encoding
UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Saviour Balzan
Affiliation: MediaToday Co. Ltd
Address: Vjal ir-Rihan, San Gwann SGN SGN 9016, Malta
Telephone: +39 0332 78-5648 or 78-9478
Fax: +39 0332 78-5154
e-mail: illum@mediatoday.com.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

3 TECHNICAL INFORMATION

3.1 *Directories and files*
1 folder containing 5,269 XML files (with one article each)

3.2 Data structure of an entry

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<article>
 <source>ILLUM</source>
 <url>http://www.illum.com.mt/2006/11/12/emmanuel_micallef.html</url>
 <date>2006/11/12</date>
 <text>
          <paragraph>
        …
          </paragraph>
…
</text>
</article>
```

3.3 Corpora size (nmb. of tokens, MB occupied on disk)
2,249,294 tokens
39.7 MB on disk

4 CONTENT INFORMATION

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

4.2 The natural language(s) of the corpus
Maltese

4.3 Domain(s)/register(s) of the corpus
News

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)

Paragraph mark-up

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

--

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

--

4.4.4 Attributes and their values (if annotated)

--

4.5 Intended application of the corpus
Text corpus to be tagged for linguistic research

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

5 RELEVANT REFERENCES AND OTHER INFORMATION