MALTESE FICTION WORDLIST

1. BASIC INFORMATION

1.1 Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),

This is a wordlist which was created from 32 Maltese fiction books. These texts were originally in PDF file format and were converted to txt format. In the next step, the text file was tokenized and a frequency count was performed on the separate tokens. The resulting list (with about 50,000 entries) was cleaned up semi-automatically.

The original list contained 46,828 tokens. After the clean-up, the list contains 41,251 tokens. The tokens were either deleted or updated (with regards to their frequencies).

Given the conversion from PDF to txt format, the list will most likely contain spelling errors that were not detected in the semi-automatic clean-up process.

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

The file is in txt format, with each line containing a token, followed by frequency (separated by comma or, in case of entries ending in hyphen or apostrophe, by six tab stops)

1.3 Character encoding

UTF-8

2. ADMINISTRATIVE INFORMATION

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

Andrew Attard (andrew.attard@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 as a zipped txt file.

2.3 Copyright statement and information on IPR

META-SHARE NonCommercial NoRedistribution licence

3. TECHNICAL INFORMATION

3.1 Directories and files

MalteseFictionFrequencyList.txt

3.2 Data structure of an entry

see 4.2

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

41,251 tokens

471 KB (469.747 Bytes) on disk

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Maltese (standard and dialect)

4.2 Entry Type

Every entry is followed by its frequency count (per million), separated by a comma, e.g.:

```
jibda,336
```

Entries in which the last character is a special character (e.g. apostrophe or hyphen) are separated from their frequency count by 6 tab stops, e.g.:

```
tal-                                    10340
ta'                                     11280
```

4.3 Attributes and their values

see 4.2

4.4 Coverage of the lexicon

Generally literal register; orate register where speech is reproduced. All in all, the books contained:

- correctly written Maltese (standard literate register)

- badly written Maltese (e.g. to mimic chat conversations)
- dialact Maltese
- English words
- Italian words
- French words

4.5 Intended application of the lexicon
For any purpose that a literature wordlist may serve…

4.6 POS assignment
n.a.

4.7 Reliability (automatically/manually constructed)
This word list is not entirely reliable, since it was converted from PDF to txt format and cleaned up only semi-automatically. It is a first version, and more refine updates should be done in the future.

5. RELEVANT REFERENCES AND OTHER INFORMATION