MALTESE SPEECH ENGINE LEXICON

1. BASIC INFORMATION

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

This lexicon is a speech lexicon, exported from Crimsonwing's text-to-speech (TTS) database into a .txt file. In its original form and together with the Maltese Speech Engine Diphone repository, it was used for building Crimsonwing's text-to-speech system.

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

The file is in txt format, with each line per word form containing the information of part of speech, written form, phonetic form, syllables, stress position and language (separated by commas).

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

MS Commons - BY - NC - SA

3. TECHNICAL INFORMATION

3.1 Directories and files

After unzipping the file, the folder contains the file

- lexicon_export_2012_12_19.txt
- Maltese_Speech_Engine_Lexicon_NarrativeDescription.doc
- Maltese TTS - Database Schema.pdf

3.2 Data structure of an entry

With one entry per line and data values separated by comma, the structure of an entry follows the structure: PartOfSpeech,WrittenForm,PhoneticForm,Syllables,StressPosition,Language

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

39,242 entries

2,1 MB (2.059.357 Bytes) on disk

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

Maltese (standard and dialect) and other languages (English, French, Italian, unknown)

4.2 Entry Type

Every line contains one word form acompanied by data values (separated by comma), according to the structure outlined in 3.2. For example, the verb form *niktbu* "we write" is represented as:

```
Verb,niktbu,nɪgdbʊ,nɪg-dbʊ,1,Maltese
```

4.3 Attributes and their values

- PartOfSpeech (Abbreviation, Acronym, Adjective, Adverb, Article, Conjunction, Interjection, Letter, Noun, Numeral, Participle, Preposition, Pronoun, Verb, Unknown)
- WrittenForm (string: orthographical representation of the entries)
- PhoneticForm (string: representation of the entries in IPA)
- Syllables (string: representation of the entries in IPA, with syllable boundaries indicated with a hyphen)
- StressPosition (number indicating the syllable carrying word stress; values are: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9)
- Language (English, French, Italian, Maltese, Unknown)

NB: The value “0” for stress position only applies to two entries which do not contain values for PhoneticForm and Syllables (for some unknown reason). The respective word forms are the adjective *iffullata* “crowded, congested” and the participle *immankat* “mutilated, disabled”.

4.4 Coverage of the lexicon

Standard Maltese (standard register as well as colloquial registers) and loan words from English, French, Italian and others.

4.5 Intended application of the lexicon

Text-to-speech systems, pronunciation dictionaries, other dictionaries

4.6 POS assignment

Abbreviation, Acronym, Adjective, Adverb, Article, Conjunction, Interjection, Letter, Noun, Numeral, Participle, Preposition, Pronoun, Verb, Unknown

4.7 Reliability (automatically/manually constructed)

This word list is automatically exported from Crimsonwing’s Maltese Speech Engine Database Lexicon.

5. RELEVANT REFERENCES AND OTHER INFORMATION

Complete documentation of the original database can be found in the accompanying documentation file

- Maltese TTS - Database Schema.pdf