F Mona 1 / Spoken Newspaper

CORPORA DOCUMENTATION

1 BASIC INFORMATION

1.1 Corpus composition

108 WAV files subdivided into 12 directories with a variable number of sentences (sometimes: clauses) each. They come together with transcriptions and tables of phoneme durations (see 1.2 below).

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

12 directories with several sentences (or clauses) each. Each sentence/clause is represented by four files:

i. A file with the speech sentence/clause - .wav

ii. A file with the sentence - .txt (This is a UTF-8 based file that uses the Maltese keyboard to include the 4 non-ASCII Maltese characters)

iii. An Excel file (xlsx, UTF-8) with 5 columns for

1. name of WAV file and beginning of phoneme (times given are in seconds)
2. end of phoneme
3. duration of phoneme
4. phoneme
5. word

1.3 Character encoding

UTF-8

2 ADMINISTRATIVE INFORMATION

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

Name: Paul Micallef
Address:
Affiliation: University of Malta, Department of Communications & Computer Engineering
Position: Professor Engineer
Telephone: +356 2340 2520
Fax: (+356) 21343577
e-mail: paul.micallef@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. (also available under: http://staff.um.edu.mt/paul.micallef/speech_annotation)

2.3 Copyright statement and information on IPR

3 TECHNICAL INFORMATION

3.1 Directories and files

12 directories, 108 WAV files, 12 txt files, 12 xlsx files

3.2 Data structure of an entry

5 columns (see above under 1.2)

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

14.3 MB zipped, 108 WAV files for 128 sentences

4 CONTENT INFORMATION

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

monolingual, audio, raw text transcriptions, phoneme alignment information

4.2 The natural language(s) of the corpus

Maltese

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

Newspaper texts

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)

n.a.

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

n.a.

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

alignment on: phoneme start, phoneme end, duration, phoneme, word

values retrieved automatically, partially manually checked

4.4.4 Attributes and their values (if annotated)

phoneme start: time
phoneme end: time
duration: time
phoneme: phoneme, sil(ence), breath, pause

4.5 Intended application of the corpus

Speech synthesis

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

automatically, partially checked

5 RELEVANT REFERENCES AND OTHER INFORMATION