CORPORA DOCUMENTATION

MalToBi/SPAN Corpus

1 BASIC INFORMATION

1.1 Corpus composition

Audio corpus: 8 subfolders with .wav files

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

1.3 Character encoding

2 ADMINISTRATIVE INFORMATION

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

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

Will be uploaded to META-SHARE - currently still to be updated/revised and available at: http://staff.um.edu.mt/cbor7/mlexweb/public_html2/maltobi/

2.3 Copyright statement and information on IPR

to be licenced under MS Commons BY-NC-SA

3 TECHNICAL INFORMATION

3.1 Directories and files

- .txt file explaining the data structure of the subdirectories and files
- 8 subdirectories, each of them containing:
  - 2 sound files containing a read story ("The sun and the wind", each by speaker A and speaker B)
  - 2 sound files containing each 30 read sentences (each by speaker A and speaker B)
  - 2 x each of the 30 sentences as a single sound file (each by speaker A and speaker B)
  - 2 x 26 phrases in individual files (each by speaker A and speaker B)

3.2 Data structure of an entry

n.a.

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

see 3.1; size on disk: 1.37 GB

4 CONTENT INFORMATION

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

Monolingual with code-switching.

4.2 The natural language(s) of the corpus
Maltese, partly English (code-switching in map tasks and dialogues)

4. 3 Domain(s)/register(s) of the corpus
Standard Maltese, literate register (for read out story, sentences and phrases), orate register (for map tasks and dialogues)

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 contain aligned documents: level of alignment, how it was achieved)
n.a.

4.4.4 Attributes and their values (if annotated)
n.a.

4.5 Intended application of the corpus
Prosody research

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

5 RELEVANT REFERENCES AND OTHER INFORMATION

MalToBi and SPAN (originally listed as two separate resources) are actually the same collection of WAV files. SPAN was the name of the project's audio corpus, while MalToBi was the name of the standard developed for transcribing Maltese intonation.