OpenProDat

Description and recording protocol

Brigitte Bigi

Corpus description

OpenProDat is an open multilingual database containing recordings of 2 texts, originally taken from the European SAM project. The database will contain both primary data, the recordings, and secondary data in the form of different annotation files. All the data will be freely available on the Speech and Language Data Repository http://sldr.org (repository number 805). The aim of this database is to collect, archive and distribute recordings and annotations of directly comparable data from a representative sample of different languages, for main speakers or learners. Linguists and engineers are welcome to download and use the corpora freely. They are kindly requested, in return, to make any additional recordings and/or annotations which they may carry out on the primary data publicly available on OpenProDat.

Recording session

1. Ask the participant to complete the information sheet document
2. Prepare recording material (see annex A)
3. Give instructions to the participant (see annex B)
4. Install participant:
    - give the texts to read ranked as follow: the mother language first, then rank from the better known language to the least known language
    - install, adjust the microphone then test sound quality
5. Perform the recording session: one recording by text to read (see annex C).

Annex A: Recording configuration

Each recording is a one channel file (mono), sampled at 48000Hz, 2 bites. Waveform Audio File Format (wav) files are primarily expected.

Annex B: Instructions to participants

“You'll have to read texts as naturally as possible. You have to read the text in black, not the information in blue. Please read texts in the order we propose. Each text will be recorded separately, so you'll have to wait our sign to start each reading.”

Annex C: The recording of one text

Steps are:

1. starts the recording
2. makes a sign to the participant to start its reading
3. stops recording
4. saves the file with the convention name (Annex D)

Annex D: File name convention

The OpenProDat file names follow a strict convention:

- 3 characters of the speaker mother language
- underscore
- 3 characters of the speaker code
- underscore
- 3 characters of the text language
- underscore
- 3 characters of the text code
- dot
- the extension

Language names are using ISO-639-3 code. Speaker code is starting by F for females, M for males and C for children, followed by a number.

Example:
FRA_F01_ENG_T02.wav is a recording of the French native speaker F01 while reading the text T02 in English.