HESITA

1. BASIC INFORMATION

1.1. Corpus composition

The **HESITA Corpus** is composed of the audio and the manual transcriptions of HESITAtion events from browdcast news in European Portuguese. The corpus includes the audio signal of 30 daily news programs collected in september 2011 from a Portuguese television channel podcast. The audio was downsampled from 44.1 kHz to 16 kHz samplig rate and the video information was discarded.

The corpus contains a total of 27 hours of audio and speech in which acoustical environment conditions and hesitations were manually transcribed by several trained annotators.

The audio material contains studio and out of studio recordings and sessions recorded from the telephone. It comprises speech (which may occur over bakground speech, noise and music) as well as non speech events (music, jingles, laughter, coughing or clapping). Prepared (read) speaking style is dominant.

For a more complete description of the corpus and the report of automatic characterization of the hesitation events, the reader may refer to (Veiga *et al.*, 2012a and 2012b), (Veiga *et al.*, 2011) and (Candeias *et al.*, 2013).

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

Two files per newsprogram are provided:

a. a WAV audio file;
b. a TRS file: containing the manual transcriptions. The TRS format is a kind of XML format that a standard transcription software such as *Transcriber* can open.
c. a TXT file: containing all the annotations presented in this corpora.

1.3. Character encoding

The characters in the text files are encoded in ISO-8859-1 (Latin1).

The audio files are uncompressed and coded in PCM WAV format.

2. ADMINISTRATIVE INFORMATION

2.1. Contact person

Name: Sara Candeias
Address: Instituto de Telecomunicações, DEEC – Universidade de Coimbra, polo II, 3030-290, Coimbra, Portugal
Affiliation: IT
Position: Post-PhD Researcher
Telephone: +351-239-796-242
Fax: +351-239-796-293

e-mail: saracandeias@co.it.pt

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 an archive.

2.3. Copyright statement and information on IPR

The resource is under CC-BY-NC-SA copyright licence.

3. TECHNICAL INFORMATION

3.1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain 58 audio files and the corresponding TRS files, that englobe the two parts of the 30 daily newsprograms.

3.2. Data structure of an entry

The TRS files have a data type definition file associated: **trans-14.dtd** that is provided in the archive.

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

TRS files have a total of 4608 hesitation events. The whole resource occupies 3GB, mainly due to the audio files.

4. CONTENT INFORMATION

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

This corpus is monolingual and annotated.

4.2. The natural language(s) of the corpus

The language of the corpus is European Portuguese.

4.3. Domain(s)/register(s) of the corpus

The corpus is comprised of RTP broadcast news.

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)

Annotations in the TRS files are at background level (temporal marker for the Repair Point), turn-level (acoustical environment, speech style and speaker) and at transcription-level with hesitation events. Such hesitation events are annotated following Shriberg's PLS (Pattern Labeling System) with some adaptations.

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

These tags are described in the DTD files. The hesitation tags are explained in the hesita_info.doc file (*to appear*).

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

Not relevant.

4.4.4. Attributes and their values (if annotated)

Described in the DTD files.

4.5. Intended application of the corpus

This corpus can be used under various purposes of research in spoken language for Portuguese in the domain of hesitation phenomenum and speech communication.

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

The annotations were done manually.

5. RELEVANT REFERENCES AND OTHER INFORMATION

(Veiga *et al.*, 2012a) Veiga, A., Celorico, D., Proença, J., Candeias, S. and Perdigão, F. 2012. "Prosodic and Phonetic Features for Speaking Styles Classification and Detection". Toledano, D.T.; Ortega, A.; Teixeira, A.; Gonzalez-Rodriguez, J.; Hernandez-Gomez, L.; San-Segundo, R.; Ramos, D. (eds.). Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science, vol. 328: 89-98. Springer (ISBN 978-3-642-35291-1).

(Veiga *et al.*, 2012b) Veiga, A., Candeias, S., Celorico, D., Proença, J., Perdigão, F. 2012. "Towards Automatic Classification of Speech Styles", in the 10th International Conference on Computational Processing of Portuguese (PROPOR 2012), April, 2012, Coimbra, Portugal. H. Caseli et al. (eds.). Lecture Notes in Artificial Intelligence (LNAI) 7243: 421-426, Springer-Verlag Berlin Heidelberg.

(Veiga *et al.*, 2011) Veiga, A., Candeias, S., Lopes, C. and Perdigão, F. 2011. "Characterization of hesitations using acoustic models", in the 17th International Congress of Phonetic Sciences (ICPhS XVII), August 17-21, 2011: 2054-2057. Hong Kong.

(Candeias *et al.*, 2013) Candeias, S., Celorico, D., Veiga, A., Lopes, C., Perdigão, F. 2013 – FCT Technical Report -- *to appear*.