LECTRA

1. BASIC INFORMATION

1. Corpus composition

This corpus is composed of the audio and the manual transcriptions of the LECTRA Corpus: classroom LECture TRAnscriptions in European Portuguese. The corpus includes seven 1-semester University courses. All lectures were taught at Technical University of Lisbon (IST), recorded in the presence of students, except IICT, recorded in another university and in a quiet office environment, targeting an Internet audience. The corpus contains a total of 21 hours of audio speech that were manually transcribed by several trained annotators.

For a complete description of the corpus and the report of Automatic Speech Recognition results, the reader may refer to (Trancoso *et al.*, 2008) and (Pellegrini *et al.*, 2012).

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

Two files per lecture are provided:
a) a RAW file: 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.

3. Character encoding

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

2. ADMINISTRATIVE INFORMATION

1. Contact person

Name: Isabel Trancoso
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: isabel.trancoso@inesc-id.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 free.

3. TECHNICAL INFORMATION

1. Directories and files

The archive that will be uploaded on the MetaShare platform will contain three folders, corresponding to three subsets of the corpus: a training, a development, and an evaluation sets.

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

The TRS files have a total of 237,984 word tokens (Training set: 189,814 word tokens, Development set: 23,714 word tokens, Test set: 24,456 word tokens). The whole resource occupies 4.2 GB.

4. CONTENT INFORMATION

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

This corpus is monolingual and annotated.

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 technical University lectures: : Production of Multimedia Contents (PMC), Economic Theory I (ETI), Linear Algebra (LA), Introduction to Informatics and Communication Techniques (IICT), Object Oriented Programming (OOP), Accounting (CONT), Graphical Interfaces (GI).

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 word-level. They are fine-grained transcriptions that include disfluencies.

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

These tags are described in the DTD files.

4.4.3Alignment 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.

5. Intended application of the corpus

This corpus can be used to train an ASR system to transcribe lectures in Portuguese.

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

The annotations were done manually. See the papers for more details about the inter-annotator agreement rates.

5 RELEVANT REFERENCES AND OTHER INFORMATION

(Trancoso et al., 2008) Isabel Trancoso, Rui Martins, Helena Moniz, Ana Isabel Mata da Silva, Maria do Céu Guerreiro Viana Ribeiro, **The LECTRA Corpus - Classroom Lecture Transcriptions in European Portuguese**, *In* LREC 2008 - Language Resources and Evaluation Conference, Marrakesh, Morocco, May 2008

(Pellegrini et al., 2012) Thomas Pellegrini, Helena Moniz, Fernando Batista, Isabel Trancoso, Ramon Fernandez Astudillo, **Extension of the LECTRA corpus: classroom LECture TRAnscriptions in European Portuguese**, *In* SPEECH AND CORPORA, Belo Horizonte, March 2012