TED Talks

1. BASIC INFORMATION

1. Corpus composition

This corpus is composed of the audio, the automatic transcriptions, the manual transcriptions and the translations for Portuguese of TED Talks from Al Gore (On averting climate crisis), Dann Dennett (On Our Consciousness) and Malcolm Gladwell (On Spaghetti Sauce). TED is the acronym for Technology, Entertainment, Design.

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

Each one of the three Ted Talk is composed of a set of files:
a) a wav file: audio file
b) a xml file: xml file containing the automatic transcription
c) a trs file: containing the manual transcriptions
d) two parallel txt files containing the manual transcription and the respective translation, without any annotation.

3. Character encoding

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

2. ADMINISTRATIVE INFORMATION

1. Contact person

Name:Luísa Coheur
Address: Rua Alves Redol, nº 9, 1000-029, Lisboa
Affiliation: IST/INESC-ID
Position: Assistant Professor
Telephone: +351 3100314
Fax: +351-213-145-843
e-mail: luisa.coheur@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, each one dedicated to one Ted Talk and being composed of the files described in Section 1.2.

2. Data structure of an entry

Both the XML and the TRS files have a data type definition file associated (respectively, alert-pt.dtd and trans-14.dtd) that are provided in the archive.

Considering the xml file, the dtd is alert-pt.dtd an example of a transcript segment is:

<TranscriptSegment>
<TranscriptGUID>7</TranscriptGUID>
<AudioType start='2421' end='2848' conf='0.456900'>Clean</AudioType>
<Time reasons='' start='2421' sns_conf='0.953300' end='2848'/>
<Speaker name='Homem' gender='M' id_conf='0.221200' gender_conf='0.921100' known='F' id='1002'/>
<SpeakerLanguage native='T'>PT</SpeakerLanguage>
<TranscriptWordList>
<Word start='2438' end='2453' conf='0.551049'>a</Word>
<Word start='2454' end='2490' conf='0.736239'>truly</Word>
<Word start='2491' end='2522' conf='0.855250'>great</Word>
<Word start='2523' end='2557' conf='0.696083'>honor</Word>
<Word start='2558' end='2563' conf='0.877075'>to</Word>
<Word start='2566' end='2588' conf='0.809444'>have</Word>
<Word start='2589' end='2602' conf='0.848018'>the</Word>
<Word start='2603' end='2670' conf='0.959818'>opportunity</Word>
<Word start='2671' end='2676' conf='0.894171'>to</Word>
<Word start='2677' end='2701' conf='0.786599'>come</Word>
<Word start='2702' end='2706' conf='0.895611'>to</Word>
<Word start='2707' end='2715' conf='0.828017'>the</Word>
<Word start='2716' end='2763' conf='0.813053'>stage</Word>
<Word start='2764' end='2809' conf='0.873642'>twice</Word>
</TranscriptWordList>
</TranscriptSegment>

Considering the trs file, an example of a turn is:

<Turn speaker="spk2" startTime="134.433" endTime="138.770">
<Sync time="134.433"/>
<this>
<Event desc="rire en fond" type="noise" extent="instantaneous"/>
This was a rented ford taurus.
</Turn>
<Turn startTime="138.770" endTime="141.304">
<Sync time="138.770"/>

<Event desc="rire en fond" type="noise" extent="instantaneous"/>
</Turn>

In the other two files each line contains a single sentence.

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

The TXT, TRS and XML files have a total of 35,399 word tokens (17,997 word tokens for the English and Portuguese TXT files). The whole resource occupies 114.9 MB.

4. CONTENT INFORMATION

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

This corpus is in multilingual, parallel and annotated.

2. The natural language(s) of the corpus

The languages of the corpus are English and Portuguese.

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

Al Gore's talk is about the climate, Dan Dannett talk is about consciousness and Malcolm Gladwell talk focus on food industry

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)

Only the xml and the trs corpora are annotated.

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 test an English recognizer, as both the audio and the manual transcriptions are provided; it can also be used to test a machine translation system for EN-PT.

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

The translation was manually done.

5 RELEVANT REFERENCES AND OTHER INFORMATION