propor 2008

International Conference on Computational Processing of Portuguese Language

Applications of Portuguese Speech and Language Technologies

Applications of Portuguese Speech and Language Technologies - Propor 2008 Special Session

Hosted by:

Universidade de Aveiro

Promoted by:

Microsoft Language Development Center

Propor 2008 Special Session Commitee

Special Session Chair

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Organising Committee

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal

Programme Committee

- **António Teixeira -** DETI/IEETA, Universidade de Aveiro, Portugal
- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Vera Strube de Lima,** Pontifícia Universidade Católica do Rio Grande do Sul, Brasil
- **Luís Caldas de Oliveira,** INESC-ID/IST, Portugal

Editorial Board

- **Daniela Braga,** Microsoft Language Development Center, Portugal
- **Miguel Sales Dias,** Microsoft Language Development Center, Portugal
- **Luanda Braga Batista,** Microsoft Language Development Center, Portugal

CORP-ORAL: a spontaneous European Portuguese speech resource

Fabíola Santos, Tiago Freitas

ILTEC , Rua Conde de Redondo, 74,1150-109 Lisboa-Portugal
{fabiola.santost, taf,}@iltec.pt

1.Overview

During the last decades there has been an increase in the development of large spoken corpora. Well-know exponents of this phenomenon are Corpus Gesproken Nederlans - CGN (Schuurman et al. 2004) and the British National Corpus - BNC (Burnard, 2002). Such corpora (and particularly subcorpora among them consisting of spontaneous speech) have been prepared to meet the growing demands deriving from different areas of language research (see eg. van Bael 2004) and speech tools development. Alongside these national enterprises there has been an emergence of supranational projects such as C-ORAL-ROM (Cresti et al. 2004) which aim to establish standardized corpora for different languages.

CORP-ORAL was created as a tool to further expand the spoken language resources available in European Portuguese (EP). Its main goal is to provide highly spontaneous interaction data with high quality audio. Developments made over existing corpora include collecting a broader range of communication contexts as well as expanding the combinations of participant ages and relations. Also, CORP-ORAL is built from scratch with the explicit goal of becoming entirely available on the internet to the scientific community and the public in general. Another relevant feature of the corpus is that it meets current metadata standards like IMDI (2003).

2.Data

Speech data consists of face-to-face dialogues between 2 participants recorded with a Marantz solid state portable recorder and Beyerdynamic microphones. Recordings are made in 44 Khz WAV format and have different lengths up to 90 minutes. Dialogues take place in partially controlled environments, mostly in closed rooms in ILTEC or in one of the informants' houses.

Each speaker is provided with an independent microphone, allowing for robust channel separation. There is some sound leaking from one channel to another but the leaked samples are never loud enough so as to prevent overlapped speech portions from being easily verified.

The data collected is spontaneous in the sense that it consists of unscripted and unprompted dialogues between family, friends, colleagues and unacquainted participants. It is not entirely spontaneous because participants are aware of the recording process and must review and sign a specific form in which they authorize proposed conditions for data collection and publication. Regardless of such constraints the recordings collected up to now show a surprising degree of spontaneity. Speakers pursue private topics and feel comfortable to adopt different speaking styles. This results in diverse voice quality settings produced by the speakers throughout the conversation, dramatic changes in pitch and tempo, etc.

Since the initial planned amount of 30 hours of spontaneous conversation has been achieved, the recording of different genres is now encouraged. Initial explorations included interaction in videogame playing contexts, and now there are ongoing experiments with commercial transaction inside shops, non-professional football match reports, tutoring sessions, etc.

Participants recorded are male and female speakers of standard EP with ages ranging from 13 to 74 years. Not all age groups are evenly represented in the corpus, the 20-30 segment being overall the best covered. An effort is being made at the final stage of the project to achieve a homogeneous distribution. There are now approximately 70 different speakers recorded, with several degrees of education ranging from 7th grade to PhD.

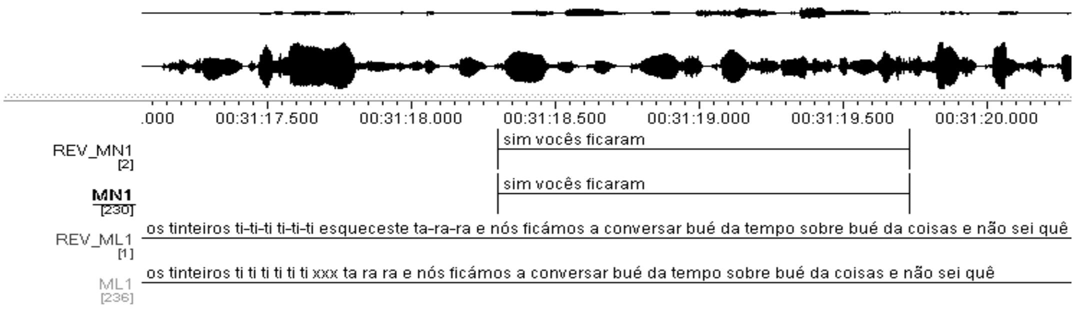
3.Annotation

3.1. Ortographic transcription

The orthographic transcription of conversations is made using ELAN software. ELAN fully supports the use of stereo files and different character sets as well as being able to export annotations to diverse file formats so they can be processed by other linguistic analysis applications such as Praat. Additionally, ELAN supports data insertion with controlled vocabularies, a particularly useful feature in tasks such as morpho-syntactic tag correction.

CORP-ORAL annotation files comprise a total of five levels primarily split in OT (orthographic transcription) and PA (prosodic annotation) tiers. The OT string is used for transcribing words in the speech and also hesitations, breathing periods, laughter and all linguistic or paralinguistic noises made with the vocal tract.

Because of the highly spontaneous quality of the material, inconsistencies are inevitable. They emerge in cases of fast and heavily repaired teenage speech such as the one presented below:



sim vocês ficaram

sim vocês ficaram

os tinteiros ti-ti-ti ti-ti-ti esqueceste ta-ra-ra e nós ficámos a conversar bué da tempo sobre bué da coisas e não sei quê

os tinteiros ti ti ti ti ti ti xxx ta ra ra e nós ficámos a conversar bué da tempo sobre bué da coisas e não sei quê

Figure 11. Orthographic transcription fragment, teenage speech, conversation #43, 31.17-31.21 seconds.

The PA string is a copy of OT without the laughter and breathing marks and with additional linguistic information regarding prosodic structure. Prosodic boundaries are inserted by the transcriber according to his or her own perception. Prosodic breaks can be of two kinds: terminal (marked with one slash) and non-terminal (marked with two slashes). This annotation system conforms to the criteria established for the previous C-ORAL-ROM project, which are published in Moneglia et al. (2005).

3.2. Phonetic Transcription

Phonetic transcription is made using Praat. This allows for the full integration of orthographic and phonetic tiers in ELAN transcription files. Three annotation tiers are created in a TextGrid file connected with the audio. The three strings consist of Seg (segment), Pal (word) and Obs (transcriber observations) tiers. In Seg, each phonetic segment is aligned with the corresponding section in the spectrum. After individual segments are identified, these are grouped as word sequences in the Pal tier. Finally, Obs is reserved for transcriber comments. Comments can be inserted relating to particular types of phonation, particular types of intonation, unusual articulations and procedural difficulties.

Although the transcriptions are made separately for mono files, they are later assembled in master TextGrid files combining a total of six tiers with the phonetic transcription of the entire conversation segment by segment. Depicted below is an excerpt of such transcription after files are joined. The visual integration of these different levels can be particularly useful for observing the continuous interaction of phonetic features in the production of talk.

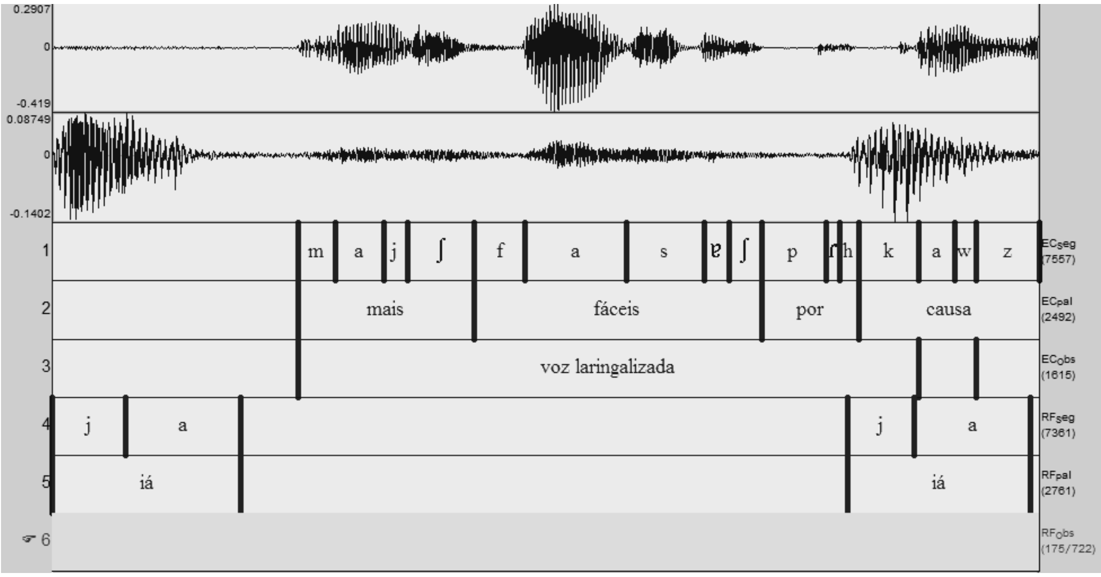


Figure 2. Phonetic transcription fragment, assembled tiers, conversation #10, 74-77 seconds.

3.3. Morpho-syntactic tagging

Morpho-syntactic tagging of all transcribed speech is currently being tested. The addition of this annotation level will greatly improve the usefulness of the corpus for linguistic research. Tests are being conducted with a version of Eric Brill tagger developed in the Linguistics Centre of Lisbon University. The manual adjustment of the automatically generated POS tags is planned in the long term using ELAN controlled vocabulary feature.

4.Results

CORP-ORAL has generated the following speech data:

	Time
Recording	60 hours
Ortographic Transcription	30 hours
Phonetic Transcription	1 hour

Table 1: Amount of speech material collected in CORP-ORAL.

5.Availability

This corpus will become available online in two different ways:

- Spock, a spoken corpus access tool
- IMDI database

Spock is a lightweight web-based application built by Maarten Janssen providing easy, online access to time-aligned corpora. It is currently in development stage. A preliminary version for browsing CORP-ORAL is accessible at:http://www.iltec.pt/spock/

During the final quarter of 2008 the corpus will also be made available in browsable full file format by accessing the IMDI database. Users interested in obtaining the corpus will be given access privileges subsequent to filling out a web-based form. The time-aligned annotation files will be downloadable in ELAN's native XML format, from which they can be converted to a number of different file types. Audio files will be downloadable in WAV format.

6.Acknowledgements

CORP-ORAL is an FCT funded project grant reference POCTI/LIN/60019/2004. Gratitude must be expressed towards CORP-ORAL consultants: Miguel Oliveira Jr., Maria do Céu Viana, Maria Fernanda Bacelar do Nascimento and Fernando Martins. Special thanks due to Maarten Janssen for the development of Spock

7.References

1. Burnard, L. (2002) Where did we go wrong? a retrospective look at the British National Corpus In Kettemann and Markus (eds.), *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi, pages 51-71.

2. IMDI (2003). *IMDI Metadata Elements for Session Descriptions*. MPI Nijmegen.

3. Moneglia, M. et al. (2005). Evaluation of consensus on the annotation of terminal and non-terminal prosodic breaks in the C-ORAL-ROM corpus. In Cresti and Moneglia (eds.), *C-ORAL-ROM: integrated reference corpora for spoken Romance languages*. Amsterdam, Benjamins, pages 257-276.

4. Schuurman, I. et al. (2004). Linguistic annotation of the Spoken Dutch Corpus: if we had to do it all over again. In *Proceedings of LREC 2004, 4th Internacional Conference on Language Resources and Evaluation*. Lisbon, Portugal: ELRA, pages 57-60.

5. van Bael, C.P.J. et al. (2004). On the Usefulness of Large Spoken Language Corpora for Linguistic Research. In *Proceedings of LREC 2004, 4th International Conference on Language Resources and Evaluation*. Lisbon, Portugal: ELRA, pages 2135-2138.