Spoken Language Corpora for Speech Recognition and Synthesis in European Portuguese

Ciro Martins†, M. Isabel Mascarenhas‡, Hugo Meinedo†, João P. Neto†
Luís Oliveira†, Carlos Ribeiro *, Isabel Trancoso†, M. Céu Viana‡
(in alphabetical order)

*†INESC / IST ‡CLUL * INESC / ISEL*
Contact person: Isabel Trancoso
Address: INESC, R. Alves Redol, 9, 1000 Lisbon, Portugal
Phone: +351 1 3100268 Fax: +351 1 3145843 E-mail: Isabel.Trancoso@inesc.pt

Abstract: *The main goal of this paper is to present the spoken language corpora collected for training and testing speech recognition and synthesis systems in European Portuguese, by the Speech Processing and Neural Networks groups of INESC, in the framework of international projects and / or joint projects with the Center of Linguistics of the University of Lisbon.*

Keywords: *speech recognition, speech synthesis, spoken language resources*

1. Introduction

According to the EAGLES handbook [3] a *spoken language corpus* is a collection of speech recordings which is accessible in computer readable form and which comes with annotation and documentation sufficient to allow re-use of the data in-house, or by scientists in other organizations. In this very general definition, the word re-use provides the motivation for this paper. In fact, the main goal of writing about the spoken language corpora we have collected is to describe them to the Portuguese language research community and thus potentially improve the reusability of these corpora.

The 5 spoken language corpora we shall describe have been collected by the Speech Processing and Neural Networks groups of INESC, in the framework of international projects and / or joint projects with the Center of Linguistics of the University of Lisbon. These corpora have been collected primarily for developing speech synthesis and recognition systems in European Portuguese. Because this development also implies the systematic study of many phenomena in phonetics and linguistics in general, some of these corpora may be (and are in fact) used for basic research.

The 5 corpora do not exhaust the list of corpora we have collected but are, by far, the most significant. Smaller corpora have been collected for specific purposes: second language acquisition, psycholinguistic research, comparison of romance languages, dialect research, etc..

In the description of the 5 corpora collected so far, we shall use a chronological order:

- EUROM.1 - read speech, 60 speakers, anechoic room, ≈ 2.6 Gb
- BDFALA - read speech, 10 speakers, sound-proof room, ≈ 2.4 Gb
- BD-PUBLICO - read speech, 120 speakers, sound-proof room, ≈ 2 Gb
- SPEECHDAT - read/spontaneous speech, 5000 speakers, telephone network, ≈ 7.5 Gb
- CORAL - spontaneous dialogue, 32 speakers, sound-proof room, ≈ 1.8 Gb

As much as possible, we shall try to describe them using the typology adopted in the EAGLES handbook, starting with the linguistic contents and the specification of the number and type of speakers, proceeding with the description of the data collection itself, and finalizing with the specification of the types of annotation included. When appropriate, details about packaging will also be included.

Before embarking on the description of each corpus per se, we shall try to pinpoint the most important categories considered in this typology, taking a few examples from the above mentioned corpora. The last section will summarize our efforts at dissemination and briefly discuss our needs in terms of spoken language resources.

2. Typology

2.1. Linguistic contents

The speech material in a corpus can vary from isolated sounds to complete conversations. On an increasing scale of "naturalness", one can distinguish several types of speech: read aloud isolated phonemes, read aloud isolated words (nonsense or existing ones), read aloud isolated sentences, read aloud text fragments, semi-spontaneous speech, spontaneous speech about a predetermined subject, Wizard of Oz and spontaneous speech.

Our 5 corpora cover the listed categories ranging from isolated words to spontaneous speech about a predetermined subject (e.g. CORAL). Missing in our list, but of particular importance for the study of human-computer interaction is the Wizard of Oz technique. In these experiments, the subject is convinced that he/she is actually talking with a computer. Hence, the human wizard which operates the "false" computer must talk with a computer voice and make deliberate errors.

2.2. Number and type of speakers

The corpus size in terms of speakers is one of the most important characteristics of a spoken language corpus. We shall consider three basic classes:

- few speakers
- many speakers (up to 50 or 60)
- very many speakers

The first class is used primarily for speech synthesis research and development, for building inventories of sub-word units and / or designing prosodic models. This type of corpus, of which BDFALA with only 10 speakers is our most obvious example, can be very large, despite being small in terms of number of speakers. A subset of EUROM.1 can also be considered a "few speakers" subcorpus.

At the other end of the scale are the "very many speakers" corpora, whose main purpose is the training and testing of speaker-independent recognition systems. Although collected in totally different environments, both BD-PUBLICO and SPEECHDAT belong to this class, with 120 and 5000 speakers, respectively.

In between the two extremes are the average-sized corpora such as CORAL (32 speakers) and the largest part of EUROM.1 (60 speakers). This type of corpus is generally developed for factorial experiments, that is, experiments in which a number of factors are defined that are hypothesized to influence some aspects of speech behavior. Hence, the number of speakers and the number of repetitions per speaker must be carefully designed.

Speaker selection is also very important and dependent on the application: for speech synthesis research (as in BDFALA), experienced speakers are most appropriate and demographic coverage is not at stake, whereas for training and testing speech recognizers, the intended user population must be carefully sampled. For the SPEECHDAT corpus, for instance, the user population covered the entire Portuguese population, excluding only very young children and very old people. Hence, a careful balance of sexes, ages and regions had to be achieved. This type of sampling is possible when speakers are collected in their own homes or offices through the telephone network. In a corpus collection such as BD-PUBLICO, on the other hand, where speakers were recorded in a studio, the number must obviously be much smaller. Therefore, the age range was restricted in order to still allow an adequate sampling of sexes and regions and to facilitate speaker recruitment in the University environment.

When selecting speakers for the EUROM.1, BDFALA and CORAL corpora, care was taken to avoid speakers with any type of speech disorders, either organic or functional, nor heavy smoking, drinking or drug habits.

The absence of professional speakers was most notoriously felt in the subset of the BDFALA corpus where emotionally loaded sentences had to be read aloud by the speakers.

2.3. Data collection

According to the typology proposed in the Eagles Handbook, data collection can be represented in 5 dimensions:

- visibility (open vs. secret)
- environment (studio vs. on location)
- control/interaction (random, spontaneous dialogue, interview and read speech recording)
- monitoring and validation (on line vs. off-line)
- data (single channel vs. multi-channel)

In all 5 corpora, the speakers volunteered for the task and were aware they were being recorded. Despite the better naturalness usually achieved with clandestine microphone recordings, difficulties of monitoring and legal problems outweigh its potential advantage.

The first corpus was recorded in an anechoic room and the following ones in a sound-proof room, except SPEECHDAT which, as mentioned before, was recorded through the telephone network on location, that is, the speakers usually called from their own homes or offices. Significative differences were observed in terms of naturalness between the anechoic room recordings and the sound-proof ones.

In terms of control, speaker prompting was used in the telephone recordings of SPEECHDAT, where the spoken prompts were played back by the automatic collection system. For the remaining read speech corpora, the text was presented either on paper or on the computer screen. In the case of the spontaneous dialogue (CORAL), the speakers were briefed before the dialogue and the conversation was unrestricted once started.

Monitoring is the task of controlling and modifying technical and phonetic characteristics on-line, i.e., during the recording process. Validation, on the other hand, relates to a posteriori evaluation of the recorded material. All our read corpora recorded on studio were monitored by an expert located outside the recording room, with audio contact with the speaker. The control of phonetic characteristics was limited to cases in which the pronunciation was not considered natural. The only exception was the read sentences of BDFALA, in which self-monitoring took place, i.e., the speaker had to listen to the sentence he/she had just recorded and could repeat it if any problem was found. Validation was used instead of monitoring in SPEECHDAT and CORAL.

In terms of data, all recordings included only the acoustic signal. Recording additional signals (such as laryngograph signals, electromyograph signal, X-rays, etc.) would have been desirable for some of the corpora (e.g. EUROM.1), but the required equipment was not available. Hence all corpora included single channel recordings, except CORAL which, in some sense, can be considered dual-channel (one for each speaker in the dialogue).

2.4. Annotation

The inclusion of some type of linguistic representation or annotation is one of the factors that determines whether a collection of speech can be considered a spoken language corpus. There are multiple levels of representation but, unfortunately, there is not a wide consensus on what should be included in each of them and how this information should be represented. Due to the fact that some of our corpora were collected in the framework of European projects, the conventions which we followed in these projects derived obviously from working groups formed during these projects.

The first level of representation which we considered is the *recording script* level. For all our corpora which included read speech (words, numbers, sentences, paragraphs), the prompting script was included.

The second level considered is the *orthographic transcription* or *transliteration* level. This level is meant to use the standard spelling conventions of the language to represent what the speakers actually said. For corpora such as EUROM.1, BDFALA and BD-PUBLICO, in which the speakers were carefully monitored, there were practically no differences between the recording script and what was actually said, so the two levels were merged into one. For corpora such as SPEECHDAT and CORAL, which were validated a posteriori, transliteration conventions had to be followed. These included how to represent reduced word forms (e.g. *tou* or *estou*), numbers, abbreviations, spelled words, filled pauses, mispronunciations, word fragments, unintelligible words, verbal deletion, etc., and also how to mark several types of noise (speaker noises such as coughs or lip smacks, stationary noises such as telephone channel noise or air condition fans, and intermittent noises such as a door slam or background speech). These problems are much more frequent in spontaneous speech, so the transliteration of the CORAL corpus was particularly hard. In addition, we had to mark turn-taking and overlap between the two speakers engaged in the dialogue.

The third level of annotation concerns linguistic levels above the phoneme such as: *morphological, syntactic, semantic and pragmatic* levels. The only corpus with both syntactically and semantically tagging will be CORAL, but this type of annotation has just been started. No other third level tagging is planned for the remaining corpora.

The fourth level of annotation is the *citation phonemic* one. This level consists of a phoneme string corresponding to the careful pronunciation of the word in isolation. This pronunciation was derived by rule, using the grapheme-to-phone conversion tool jointly developed by CLUL and INESC, and manually corrected a posteriori. The SAMPA phonetic alphabet (Speech Assessment Methods Phonetic Alphabet) was adopted in all cases. A pronunciation lexicon is included in both BDFALA, BD-PUBLICO and SPEECHDAT.

The fifth level of annotation is the *broad phonetic* or *phonotypic* one. This level may be derived from the previous one by phonological rules and uses symbols that have the same status as phonemes, marking the output of connected speech processes that either insert or delete phonemes or transform one phoneme into another. The EUROM.1 corpus includes this type of representation, also using SAMPA. However, it is worth noticing that some authors claim that phonotypic transcriptions can only be made by listening to the speech signal.

The sixth level of annotation is the *narrow phonetic* one. For this type of representation, manual inspection of the speech signal is mandatory. The inventory of symbols must be augmented to include sounds with no phonemic status, and mark different allophones, devoicing or voicing, nasalization, labialization, etc.. However, one segment at this level can correspond to more than one portion of speech that is recognizably separate when observing the acoustic waveform or the spectrogram (e.g. a voiceless plosive can be separated into closure plus burst). The next level, designated as *acoustic-phonetic* includes this fine distinction. The acoustic-phonetic annotation of a subset of EUROM.1 is currently in progress.

The eighth level is the *physical* level. This level is mostly used when multi-channel recordings are available and separate tiers related to the different inputs (e.g. nasal transmission detectors, palatography, etc.) must be adopted. Therefore, this detailed level was not used in our corpora.

The final level is the *prosodic* one. Absolute recommendations concerning this level are not available yet, as there is not a convincing amount of evidence concerning the adequacy of the existing labelling systems for a variety of purposes in a multilanguage framework. As very little is known about European Portuguese prosody and the way it interfaces with morphology, syntax and semantics, an intonational approach of phrasing is currently being used, which follows most of the ToBI (Tone and Break Indices) conventions ([9], [1]). The prosodic labelling splits, thus, into several tears, each one marking a different type of event. Besides the orthographic word tear, a Break Index tear with 4 degrees of cohesion/disrupture between consecutive words and a Tone tier with pitch accents and boundary tones, are provided. At the tone tier, however, extensive use is made of ToBI diacritics to indicate starting and ending times of tonal events, as well as pitch range.

Given the importance of disfluency patterns for a better understanding of the underlying speech production mechanisms and dialogue structuring, all phenomena at this level are annotated in a fourth separate tear (eg. silent and filled pauses, word, subword or part of speech suppression, truncation, contraction, etc.). The annotation conventions proposed in [8] are currently being used at this level. The ToBI miscellaneous tear is kept for other annotations. Different parenthesis are used, however, to represent human vocal noises related to the communication situation (laughs, coughs, etc.) and other types of noises (e.g. microphone, chair). The criteria for prosodic annotation, more extensively described in [10], are currently being tested on spontaneous speech in the framework of the CORAL project.

In the above discussion, we have talked about annotation symbols or labels but not about segmentation, although the two processes are closely interlinked. Segmentation can be done manually, automatically or semi-automatically. Manual segmentation is extremely costly in time and effort. However, if the string of labels is available together with acoustic models for each of these labels, alignment programs based on the Viterbi algorithm can be used to automatically align these labels with the speech signal. Manual correction of these time stamps is far easier than starting from scratch. Moreover, the segmented corpus can be used to derive better models for each label and thus contribute to improve the aligner itself in a bootstrap procedure. A subset of EUROM.1 was segmented semi-automatically in this way. This segmented corpus has been used to derive initial recognition models for the remaining corpora. The format of the label files is the one used by the WAVES program.

3. EUROM.1

The EUROM.1 corpus for European Portuguese [2] was collected in the framework of the SAM_A European project, jointly by INESC and CLUL. This project was in fact an extension of a preliminary project (SAM - Speech Assessment Methods) during which work on the planning of a poly-language resource for the Spoken Language Engineering needs of the European Union was first started. Despite its main use for recognition and synthesis research, this courpus has also been extensively used in our group for phonetic coding research.

3.1. Linguistic contents

For each of the 11 languages contemplated in this project, 4 types of corpus material were collected:

- CVC material (totalling 121 different logatomes) in isolation and in context (5 carrier phrases)
- 100 selected numbers from 0-9999
- 40 short passages each containing 5 thematically connected sentences (half of the passages were freely translated from the English version of EUROM.1; most of the remaining ones were adapted from Portuguese books and newspapers)
- 50 filler sentences to compensate for the phoneme-frequency imbalance in the passages

3.2. Number and type of speakers

The corpus was structured into 3 target corpora subsets:

- Many Talker Corpus (30 male + 30 female speakers): 100 numbers, 3 passages, 5 sentences
- Few Talker Corpus (5 male + 5 female speakers, selected from MANY): 5 x CVC material, 5 x 100 numbers, 15 passages and 25 sentences.
- Very Few Talker Corpus (1 male + 1 female speakers selected from FEW): CVC in context

The speakers were selected to cover a wide range of age groups and normal voice types. One main accent group was selected (Lisbon area), together with a small number of speakers from other accent regions.

3.3. Data collection

The recordings were made in an anechoic chamber environment using a high quality microphone, directly to disc (using an A/D board), and also to DAT tape. The EUROPEC program was adopted, prompting the items to be read on the computer screen). The sampling frequency defined for the project was 20 kHz. The calibration procedure followed the SAM recommendations as well. Careful monitoring was adopted.

3.4. Annotation

The SAM project defined the format of the label files which were produced. Besides the orthographic transcription, these included information about the signal file and the recording session, among other items.

3.5. Packaging

The corpus is contained into 5 CDROMs and totals 2.6 Gb.

4. BDFALA

The BDFALA corpus [10] was jointly developed by INESC and CLUL in the framework of the national project with the same name, sponsored by JNICT (Program Lusitânia). The project had as its main goal the enlargement of the EUROM.1 corpus for European Portuguese, mainly for the improvement of speech synthesis systems in our language.

4.1. Linguistic contents

6 types of corpus material were collected:

- ≈ 4600 isolated words
- 350 sentences for prosodic studies
- 18 phonetically-complete paragraphs
- 60 read paragraphs extracted from television debates
- ≈ 3000 logatomes
- 600 phonetically rich sentences

4.2. Number and type of speakers

The 8 speakers were selected to achieve a balance in terms of sex (4 male + 4 female), age groups (between 20 and 50) and, as much as possible, among speakers of the EUROM.1 corpus. The two latter corpus types were only spoken by one male and one female speakers. A subset of the corpus was also read by two young speakers (one male and one female), 12-14 years old, which were also recorded in the SAM_A project.

4.3. Data collection

Data collection took place in a sound-proof room. Two recording modes were adopted: in the case of isolated words and logatomes, the material was read from paper and recorded directly to DAT. Repetitions were made whenever the speaker or the person monitoring the recording decided to. The speech material was semi-automatically segmented into words and validated a posteriori. In the second mode used for the remaining sentences and paragraphs, a self-monitoring program was adopted which recorded directly into disc. After reading each sentence or paragraph, the speaker had the alternative of listening to the recorded material, recording it again or proceeding to the next item to be read. The recordings were duly calibrated in both cases. The sampling frequency was 16 kHz.

4.4. Annotation

For each spoken item, the corresponding orthographic script is saved in a separate ASCII file. A pronunciation lexicon with citation phonemic transcriptions for each word is also included. These were automatically produced and hand-corrected a posteriori.

4.5. Packaging

The corpus material amounts to around 2.4 Gb, and is stored in 4 CDROMs.

5. BD-PUBLICO

The BD-PUBLICO [5] database (Base de Dados em Português eUropeu, vocaBulário Largo, Independente do orador e fala COntínua) was collected by INESC in the framework of a national project (PRAXIS XXI Program), an European project (SPRACH) and with the collaboration of Instituto Superior Técnico (IST) and the PÚBLICO newspaper.

In this database we impose a dictation task, but in a speaker-independent mode, given the number of speakers and the quantity of data for each speaker. It might not be the best situation in terms of continous speech recognition but it will help us to develop domain independent acoustic models, pronunciation dictionaries and language models and, thereby, domain independent continuous speech recognition systems. Our aim was to create a corpus equivalent in size to the WSJ0 database [6].

5.1. Linguistic contents

For this purpose, we collected a large corpus of newspaper text extracted from the Portuguese newspaper PÚBLICO (a daily newspaper, with a broad coverage of subjects, writing styles and available on the WEB). We collected 6 months of news, resulting in a total of 10M words and 156k different words.

The newspaper text was divided into three parts: training (80%), development (10%) and evaluation (10%). This selection was made in a random fashion having the paragraph as unit.

Speakers were asked to read a set of sentences extracted in paragraph blocks from this text. For both the development and evaluation test sets we decided to have two vocabularies: a small one with no more than 5K words and a larger one with no more than 20K words.

The overall selected sentences were individually examined, to eliminate those that were hard to read. Then they were converted into prompts to be used in the recording phase, and into standard SGML format to be used in the recognizer score.

Additionally, 15 speaker-adaptation and 3 calibration sentences were selected. These sentences were chosen to be phonetically rich. They were originally from the PÚBLICO texts but were modified by hand.

The numbers of sentences are the following for each set:

- Training set with 80 sentences plus 3 calibration sentences for each speaker.
- Test sets with 40 sentences plus 15 speaker-adaptation sentences and 3 calibration sentences for each speaker.

The allocation of the sentences to the speakers was random, with sentence replacement between speakers.

5.2. Number and type of speakers

Speaker selection was done among undergraduate and graduate students from IST, a large engineering school of the Technical University of Lisbon. Ages ranged between 19 and 28 and a broad coverage of accents was obtained.

We recorded a total of 120 speakers with 100 for the training set (50 male and 50 female) and 20 speakers (10 male and 10 female) divided equally in the 5K word sets (evaluation / development). Each recording session resulted in 15 minutes of speech.

We expect to record the 20K words sets in a later phase. It will be simple to create new test sets because the recording conditions and the students will still be available.

5.3. Data collection

The recordings were done in a sound proof room at INESC (Lisbon) using a high quality microphone, directly to disc with 16kHz sampling frequency.

5.4. Annotation

A pronunciation lexicon with citation phonemic transcriptions for each word was produced by hand-correcting the automatically generated transcriptions.

5.5. Packaging

The corpus material amounts to more than 2 Gb, and we expect to release it in 4 CDROMs.

6. SPEECHDAT

The SPEECHDAT [4] corpus collection for European Portuguese was divided into 2 phases: the first phase comprised the collection of 1000 telephone calls and was done in the scope of the preparatory MLAP Program. The second phase comprises 4000 telephone calls and is taking place in the framework of the Language Engineering project with the same name, which incorporates databases from all official languages of the European Union and some major dialectal variants. In both projects, the work is done by INESC under a subcontract with Portugal Telecom. The second phase is not yet complete ($\approx$ 1200 missing).

The main goal of these databases is to provide a realistic basis for training and assessment of both isolated and continuous speech utterances, either using whole word or subword approaches, and thus can be used for developing voice driven teleservices.

6.1. Linguistic contents

In the current phase, each speaker is asked to answer a few spontaneous questions (7), some of them related to demographic information (e.g. date and place of birth) and to read a prompt sheet with 33 items. 4000 different prompt sheets were produced. The recording material for each speaker comprises:

- 3 application words (chosen from a vocabulary of 30)
- 1 sequence of isolated digits
- 4 connected digits (prompt sheet, telephone, credit card, PIN code)
- 3 dates (1 spontaneous (date of birth) + 2 read)
- 1 word spotting phrase
- 1 isolated digit
- 1 currency amount
- 1 natural number
- 3 spelled words(1 spontaneous (forename) + 2 read)
- 5 directory assistance names (2 spontaneous (forename and place of birth) + 3 read)
- 2 questions (predominantly yes and no, but also fuzzy answers)
- 2 time phrases (1 spontaneous (time of day) + 1 read)
- 4 phonetically rich words (chosen from a set of 4000)
- 9 phonetically rich sentences (chosen from a set of 3600)

6.2. Number and type of speakers

Speaker selection is done among employees of Portugal Telecom and their relatives and friends. In the current second phase, the demographic coverage is as follows: Entre-Douro-e-Minho - 31%, Estremadura - 23%, Beira-Litoral - 14%, Alentejo - 7%, Trás-os-Montes - 5%, remaining regions - less than 5% each. Although not required in the project specifications, we are trying to reach a minimum of 5% for the remaining regions.

The age distribution exceeds 20% for the 3 main age groups considered: 16-30, 31-45 46-60. Gender distribution is close to ideal (47% male and 53% female).

6.3. Data collection

The design of the collection platform and the speech data collection itself are the responsibility of INESCTEL. The recording platform consists of a PC equipped with two Dialogic boards: the DTI/212 board connects the 8 channels of the D/81A board to the E-1 digital network. The speech signals are stored at 8kHz, 8-bit A-law format.

6.4. Annotation

Each speech file has an accompanying ASCII SAM label file. These files include information about the calling session, the recording conditions, speaker sex, age and accent, signal file, recording date and time, assessment codes and the label file body itself. This includes both the prompting script and the orthographic transcription. A pronunciation lexicon with citation phonemic transcriptions for each word is also included. These were automatically produced and hand-corrected a posteriori.

6.5. Packaging

The corpus material for the first phase (1000 speakers) is stored in 3 CDROMs. The phonetically rich sentences are contained in one CDROM and the remaining material in two. All the signal files were compressed via *GNU zip*. The first 1000 speakers of the second phase are stored in 3 CDROMs, but the signal files were not compressed.

7. CORAL

The CORAL corpus [11] is being collected in the framework of a national project sponsored by the PRAXIS XXI program, by a consortium formed by INESC, CLUL, FLUL (Faculdade de Letras da Universidade de Lisboa), and FCSH-UNL (Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa).

The purpose of this project is the collection of a spoken dialogue corpus, with several levels of labelling: orthographic, phonetic, phonological, syntactic and semantic.

7.1. Linguistic contents

The task of each dialogue is to follow map directions, as done already in several other languages. One of the participants in the dialogue has a map with some landmarks and a route drawn between them; the other has also landmarks, but no route and consequently must reconstruct it. In order to elicit conversation, there are small differences between the two maps: one of the landmarks is duplicated in one map and single in the other; some landmarks are only present in one of the maps; and some have slightly different names in the two maps (e.g. *curvas perigosas* vs. *troço sinuoso*).

In the 16 different maps, the names of the landmarks were chosen to allow the study of some connected speech phenomena. The following contexts were elicited:

- sequences with /l/ favouring or not its velarization (e.g. *sala malva, sal amargo*)
- Sequences with /s/ in word final position followed by another coronal fricative (e.g. *barcos salva-vidas*)
- Sequences of plosives formed across word boundaries (e.g. *clube de tiro*)
- Sequences of obstruents formed within and across word boundaries (e.g. *bairros degradados*)

The last three items were designed to allow a more comprehensive study of consonant clusters formed within and across word boundaries and should, therefore, be jointly investigated.

7.2. Number and type of speakers

The corpus includes 32 speakers, which were divided into 8 quartets and, in each quartet, organized to take part in 8 different dialogues. Given the reduced number of speakers, they were chosen to achieve an adequate balance of sexes, but were restricted in terms of age (under-graduate or graduate students) and accent (Lisbon area). Speakers are chosen in pairs who know each other, so that half of the conversations take place between "friends" and half between people who do not know each other.

7.3. Data collection

The recordings take place in a sound proof room at INESC, with no visual contact between the speakers engaged in the dialogue. The speakers wear close-talking microphones and the recordings are made in stereo directly to DAT and later down-sampled to 16 kHz per channel. Before each conversation, the speakers are briefed and recording levels are adjusted. No monitoring is done once the recordings start. The recording phase is not yet complete.

7.4. Annotation

As mentioned above, this corpus will be annotated in several levels. For the recordings done up to the moment, only orthographic transcription was included. Due to severe funding restrictions, only a subset of the corpus will be annotated at all levels.

8. Dissemination and Future research efforts

All these databases will be available in the near future through the ELRA Agency, although the distribution of the SPEECHDAT corpus is the responsibility of Portugal Telecom.

A considerable effort is still needed in order to provide a full and coherent annotation of all the presented spoken corpora. This on-going effort is crucial for a multiplicity of purposes, including basic linguistic research as well as the development of different types of applications for European Portuguese.

The needs for further spoken language resources in European Portuguese are not exhausted with these corpora. We have already mentioned the need for Wizard of Oz collections, for instance, and in large vocabulary speech recognition, topic spotting and speech synthesis research, there is also a need for a corpus which incorporates a variety of speaking styles and channel conditions. Broadcast news from both television and radio programmes provide a good framework for this type of research. This collection is planned for the near future and, if possible, we would like to do it in the framework of an international project, so that corpora of comparable structure and dimension could be collected, in order to enable the evaluation of future results.

References

[1] M. Beckman and G. Ayers Elam, *Guide to ToBI Labelling*, Electronic text and accompanying audio example files available at http://ling.ohio-state.edu/Phonetics/E_ToBI/etobi_homepage-.html, 1994/1197.

[2] D. Chan, A. Fourcin, D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld and J. Zeiliger (in alphabetical order), *EUROM - a Spoken Language Resource for the EU*, Proc. European Conference on Speech Technology, Madrid, September 1995.

[3] Dafydd Gibbon, Roger Moore and Richard Winski (Editors), *Handbook of Standards and Resources for Spoken Language Systems*, 1997.

[4] H. Höge, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach and K. Choukri, *European Speech Databases for Telephone Applications*, Proc. Int. Conf. Acoust., Speech and Signal Processing, Munich, March 1997.

[5] J. Neto, C. Martins, H. Meinedo and L. Almeida, *The Design of a Large Vocabulary Speech Corpus for Portuguese*, Proc. European Conference on Speech Technology, Rhodes, Greece, September 1997.

[6] D. Paul and J. Baker, *The Design for the Wall Street Journal-based CSR Corpus*, Proc. International Conference on Spoken Language Processing, Banff, Alberta, pp. 899-902, 1992.

[7] J. Pitrelli, M. Beckman and J. Hirschberg, *Evaluation of prosodic transcription labelling reliability in the ToBI framework* Proc. International Conference on Spoken Language Processing, Yokohama, Japan, Vol. 1, pp. 123-126, 1994.

[8] E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*, PhD Diss., University of California at Berkeley, 1994.

[9] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Whightman, P. Price, J. Pierrehumbert and J. Hirschberg, *ToBI: a standard for labeling English prosody*, Proc. International Conference on Spoken Language Processing, Banff, Alberta, pp.867-870, 1992.

[10] I. Trancoso, C. Viana, L. Oliveira, M. Mascarenhas, P. Carvalho and C. Ribeiro, *BDFALA Final Report*, June 1997.

[11] C. Viana and I. Trancoso, *CORAL Internal Report (Ref. CORAL/97/08)*, November 1997.