Spoken Portuguese: Geographic and Social Varieties

José Bettencourt Gonçalves, Rita Veloso

Centro de Linguística da Universidade de Lisboa Av. 5 de Outubro, 85 - 6°, 1050-050 LISBOA, Portugal {jose.bettencourt, rita.veloso}@clul.ul.pt

Abstract

The Spoken Portuguese: Geographic and Social Varieties project has as its main goal the Portuguese teaching as foreign language. The idea is to provide a collection of authentic spoken texts and to make it friendly usable. Therefore, a selection of spontaneous oral data was made, using either already compiled material or material recorded for this purpose. The final corpus constitution resulted in a representative sample that includes European, Brazilian and African Portuguese, as well as Macau and East-Timor Portuguese. In order to accomplish a functional product the Linguistics Center of Lisbon University developed a sound/text alignment software. The final result is a CD-ROM collection that contains 83 text files, 83 sound files and 83 files produced by the sound/text alignment tool. This independence between sound and text files allows the CD-ROM user to manipulate it for other purposes than the educational one.

1. Introduction

This project was carried out by the Linguistics Center of Lisbon University (Centro de Linguística da Universidade de Lisboa) and was sponsored by Instituto Camões and by the European Program LINGUA-SOCRATES, which aims to promote the knowledge of foreign languages in the European Community, within the ACTION VB – Development and Interchange of Teaching Materials. The Linguistic Center of Lisbon University is the co-ordinator institution of the project, with partnership of the Universities of Toulouse-le-Mirail and Aix-en-Provence, France. The works are already finished and its publication in CD-ROM is foreseen for the next months. The recordings are samples of oral Portuguese, formal and informal, covering Portugal, Brazil, the different African countries with Portuguese as its official language, Macau and East-Timor, as one can see in Figure 1, in a total of 83 recordings, corresponding to nearly eight hours and thirty minutes speech.

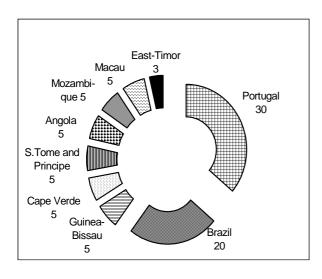


Figure 1: Documents Origin

The documents selection covers a three decades period - seventies, eighties and nineties - although the last decade is the most represented one. (See Figure 2).

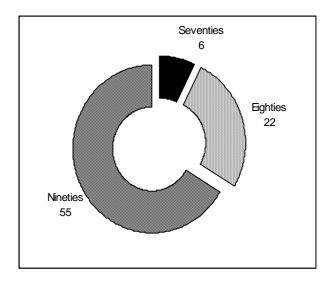


Figure 2: Documents Decade Distribution

The CD-ROMs contain the sound files, their correspondent orthographic transcription in text files and an application developed in the institution which aligns the sound with the text: a colored light runs over the transcription of the sequence which is being listened to. The user can control what he is listening to, repeating sequences or jumping parts of the text.

2. The Main Goal of this Project

Portuguese is one of the less taught languages in the European Union in spite of being the third European language spoken in the world, reason why this project can be truly important for the improvement of its teaching. As a matter of fact, in the last two decades, the importance of the use of authentic documents and of the oral in teaching of foreign languages has been growing. As far as oral is concerned, a kind of prejudice has led teachers to look at it as something imperfect, a kind of denial of the grammar rules. Since everybody realizes that listening and producing oral messages is crucial, the forgery of oral texts became a common practice: texts are read by

professional actors with an ideal pronunciation, and the apparent irregularities and particularities of the oral speech are deleted. As a consequence of this spirit, the students who learn a foreign language in their own countries, many times are used to the way of speaking of their teachers and, eventually, to the "artificial" pronunciation of these texts read by professionals. When they are confronted with a real situation of communication in the language they are learning they often feel lost. This happens because what they are listening to doesn't match with what they are used to listening to: either their interlocutors speak too fast, or the communication strategies of the interlocutor are very different from those acquired by the learner in the classroom.

The goal of this project is then to provide authentic speech texts representing different varieties of the Portuguese language. Each text has a sound component (in a wave file) and a written one (in a text file). For the transcription the orthographic representation was chosen (criteria being always controversial and possible solutions always showing advantages and disadvantages) considering the benefits it can bring to students who are usually used to the language orthographic representation. The student can listen to real speech situations without feeling frustrated: now he can listen and read at the same time, having the orthographic support of any misunderstood part of the speech as well as the text/sound timing control.

3. The Corpus Constitution

Having in mind that no language, although its unity, is uniform, this collection of samples covers a large range of regional, social and situational realizations of the Portuguese language. Varieties from different countries having Portuguese as its official language were selected, as well as different dialects within the Portuguese and the Brazilian territory. The informants were selected from different levels of education, different professional status and covering a wide range of ages. As far as European and Brazilian Portuguese is concerned, also diachronic variation was taken into account.

3.1. Documents Selection

Material selection involved many different factors. At the project beginning, a large amount of documents was selected from the oral *sub-corpus* of the *Contemporary Portuguese Reference Corpus*, including some material provided from Brazilian and Mozambican *corpora*. However, further more variety samples had to be specifically collected for the project.

The first criterion considered in a pre-selection was the sound quality of the material. This was very restrictive, since there were some documents recorded in the seventies and eighties, being its sound quality very poor. After that, the selection related mostly with the language variety representation, considering dialects and sociolinguistic factors, such as age, gender and educational level, as referred above. The figures 3, 4, 5 and 6 show the final selected data distribution according to these variables. The educational level was divided in three categories: until 6 years of scholarship; from 7 to 12 years of scholarship; more than 12 years of scholarship.

People were also divided in three groups according to their age: from 15 to 30 years old; from 31 to 45 years old; and more than 46 years old.

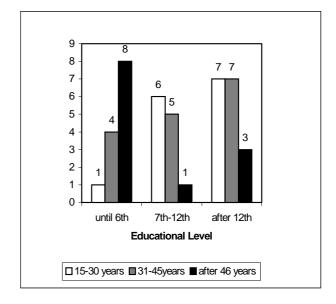


Figure 3: Women Age / Educational Level Distribution

The data shown is these figures only refer to the informants about which there was available information. Some documents were collected from radio interviews and it was not always possible to determine the precise data. It is also worth to mention that the number of informants does not exactly correlate with the number of documents, since in the radio interviews there are sometimes more than one informant.

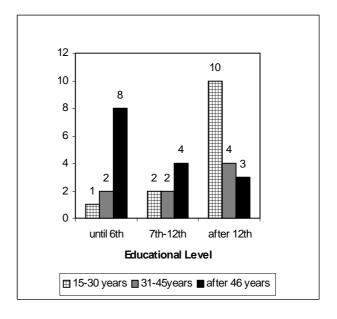


Figure 4: Men Age / Educational Level Distribution

As one can see in the charts, the sociolinguistic variables were not always balanced, regarding the project didactic aims. The topic appeal and diversity was the final used criterion, as well as the speech clearness, leading sometimes to a new evaluation of non-selected material.

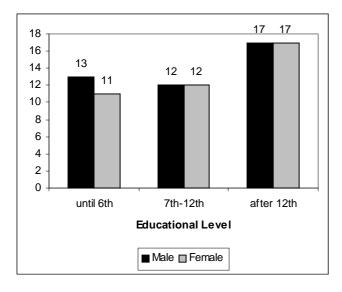


Figure 5: Educational Level / Gender Distribution

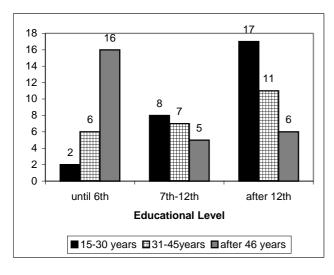


Figure 6: Educational Level /Age Distribution

3.2. The Final Result

The *Corpus* ended with 83 texts, corresponding to 30 Portuguese documents - about 3h speech; 20 Brazilian documents - about 2h speech; 25 African documents - about 2h 50m speech; 5 documents from Macau - about 38m speech; and 3 documents from East-Timor - about 10m speech -, in a total of nearly 8h30m speech.

4. The Alignment Task

A software application had to be designed in the institution for the alignment task. This tool allows the association between the text image and the sound wave. This association is established relating groups of characters with time intervals. The program opens the transcription and the sound files. When the operator starts playing the sound he can click on the text part he wants to associate with the correspondent speech part. Some adjustments can be done manually.

The alignment task considered three different types of units - syntactic, prosodic and rhythmic - which had to be coherently combined in a way that would allow a clear text/sound relation, i. e., whenever the prominent unit sounds, the correspondent text part highlights. The type of unit chosen in each case depended on the length of graphical part of text corresponding to the speech interval, in order to avoid delays between sound and highlighting, minding the student as a final user.

5. CD-ROMs Content

In 4 CD-ROMs the user can find use instructions and the project and material description. The documents are divided in folders according to its origin. Each document have 3 related files: a text file, in a txt format, with the transcription preceded by a heading containing the interview specific data – title, origin, year of collection – and informant specifications – sex, age, level of education, professional status – as well as different kinds of observations, as some expressions particular use or recording situation details; a sound file (wav format), a text/sound alignment file generated by the software tool described above (dat file). Another software application is available, this one allowing the user to listen and read the selected document at the same time.

5.1. The Lingua Tool

This friendly application, designed for a Windows environment and also developped in the institution, permits the user to open a selected document and to manipulate it while listening to it. With the familiar buttons of a tape recorder, it is quite easy to work with. The user has before him the text image and in the toolbar there are a play button to start the sound playing as well as the respective transcription highlighting. The sound can be controlled through a pause button for temporary suspensions, a stop button to finish the sound playing and rewind and forward buttons to repeat or jump parts of the sound playing – these last tasks can also be accomplished by mouse clicking in the desired part of the transcription.

6. Other Applications

Due to its diversity, this *corpus* constitutes a very useful tool not only for training the capacity of listening and understanding in the teaching activity and for textual analysis, but also for different works on the Portuguese language.

For its characteristics, it allows the user to choose the texts according to its needs and preferences, fact that will give him a considerable degree of autonomy in his activity. Once the materials were not selected having in mind a restricted profile of public, besides its usefulness in the process of Portuguese teaching as a foreign language, they can equally be of great interest for the training of translator-interpreters, as well as in first language teaching. Due to the autonomy of the text and sound files, the 83 documents can also be explored in research projects on spoken Portuguese. The sound files can be used in phonetic and dialectology research and the text files can be manipulated as a useful corpus for spoken Portuguese. Regarding its constitution, it is a good sample of a wide set of Portuguese varieties. It would then constitute a reliable source of information, allowing the extraction of different kinds of data, such as concordances, lexical and syntactic associations and frequencies. For all these characteristics, this project will be an original contribution for the knowledge of spoken Portuguese

7. References

- Abaurre, M.B.M. e Â.C.S. Rodrigues (orgs.), *Gramática do Português Falado*, vol. VIII. Campinas: Editora Unicamp (in edit.).
- Bacelar do Nascimento, M.F. 1996a. Aspectos da sintaxe do português falado (repetições lexicais e de estruturas sintácticas em produções orais: fenómenos de deslocação). In I. Duarte e I. Leiria (orgs.), *Actas do Congresso Internacional sobre o Português*, Lisboa: APL, I: 203-223.
- Bacelar do Nascimento, M.F. 1996b. A observação e análise de dados reais na investigação e ensino de línguas", Actas do II Encontro da Associação Portuguesa dos Centros de Línguas do Ensino Superior, Universidade de Évora (in edit.).
- Bacelar do Nascimento M.F. 1999. Exploração de dados naturais na aprendizagem do português. In Workshop *Uso de corpora linguísticos na educação*, Maputo (in edit)
- Bacelar do Nascimento, M.F., M.L Garcia Marques and M.L. Segura da Cruz 1987. *Português Fundamental*, Vol.2 *Métodos e Documentos*, Tomo I *Inquérito de Frequência*, Lisboa, INIC/CLUL.
- Bacelar do Nascimento, M.F., P. Rivenc and M.L Segura da Cruz 1987. Português Fundamental, Vol.2 -Métodos e Documentos, Tomo II Inquérito de Disponibilidade, Lisboa, INIC/CLUL.
- Bacelar do Nascimento, M.F., M.C. Rodrigues e J. Bettencourt Gonçalves (orgs.) 1996 *Corpora, Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, Lisboa: APL.
- Blanche-Benveniste, Cl. and C. Jeanjean 1987. Le Français Parlé, Transcription et Edition, Paris, INALF.
- Stroud C. e P. Gonçalves (orgs.) 1998. Panorama do Português oral de Maputo Vol. I Objectivos e Métodos, Cadernos de Pesquisa nº 22 INDE, 1998, Vol. II A Construção de um Banco de Erros, nº 24, INDE, 1998, Vol. III Estruturas Gramaticais do Português: Problemas e Exercícios, nº 27, INDE.

Viana, M.C., I. Trancoso, I. Mascarenhas, L.C.Oliveira e C.M. Ribeiro 1996. Corpora de Fala em P.E., Constituição, Segmentação e Etiquetagem. In Bacelar do Nascimento, M.F. et alii (orgs.), 189-216.