PROVIDING ON-LINE ACCESS TO PORTUGUESE LANGUAGE RESOURCES: CORPORA AND LEXICONS

Maria Fernanda Bacelar do Nascimento, Amália Mendes, Luísa Pereira
Centro de Linguística da Universidade de Lisboa (CLUL)
Av. Prof. Gama Pinto, 2 - 1649-003 LISBOA
{fbacelar.nascimento, amalia.mendes, luisa.alice}@clul.ul.pt

ABSTRACT

Several Language Resources (LRs) for Portuguese, developed at the Center of Linguistics of the Lisbon University (CLUL), are available on-line at CLUL’s webpage: www.clul.ul.pt/english/sectores/projecto_rld.html. These LRs have been extracted from or developed based on the Reference Corpus of Contemporary Portuguese (CRPC’), a monitor corpus containing, at the present, more than 350 million words, taken by sampling from several types of written text (literary, newspaper, technical, didactic, juridical, parliamentary, etc.) and spoken text (informal and formal), pertaining to national and regional varieties of Portuguese (including European, Brazilian, African and Asian Portuguese).

The LRs available for on-line queries include: a) several subcorpora (written and spoken, tagged and untagged) compiled and extracted from CRPC for specific CLUL’s projects and now available for on-line queries; b) a published sample of “Português Fundamental”, a spoken CRPC subcorpus, available for texts download; c) a frequency lexicon extracted from a CRPC subcorpus available for both on-line queries and download. Other RLs available for Portuguese are also referred: C-ORAL-ROM - Integrated Reference Corpora for Spoken Romance Languages, a CD-ROM edition of a spoken corpus with text-to-sound alignment; the LE-PAROLE corpus; the LE-PAROLE Lexicon and the SIMPLE Lexicon.

ON-LINE QUERIES TO EUROPEAN PORTUGUESE CORPORA

Some CRPC subcorpora have been developed under specific projects at CLUL. These resources are available for on-line queries at CLUL’s webpage, using CLUL’s concordancer CONCOR adapted to run on the Internet, as well as CLUL’s lemmatiser. When searching for a lemma or a wordform, it is possible to choose the corpus, to ask for concordances or frequencies, to sort the concordances results, to establish the context length and to obtain bibliographic references.

The corpora available for on-line queries are the following:

CORPUS RL

This corpus was compiled under the project Language Resources for Portuguese (Program Lusitânia and Fundação Calouste Gulbenkian). The corpus has 9,171,480 words, with a tagged subpart (See distribution in Table 1), and is available at:

The Language Resources for Portuguese non-tagged subcorpus is composed of 8,5 million words, both written and spoken European Portuguese discourse, selected from CLUL’s Reference Corpus of Contemporary Portuguese.

The written discourse texts are extracted from books, newspapers and magazines, and also from a miscellaneous of leaflets, brochures, official documents, etc. These texts are relating to literary, informative, scientific, technical and didactic genres, in a wide diversity of domains. The spoken subcorpus is composed of 105964 running-words as described below (See Português Fundamental subcorpus).

- The Language Resources for Portuguese morphosyntactically tagged subcorpus has about half million words and is composed of written discourse from newspaper, literary book, technical periodical and other (varia) texts samples.

- The objective was to achieve a large annotation, with the basic morphosyntactic categories, and to reduce ambiguities in order to arrive at the minimum possible error rate, when using the automatic tagger (http://www.clul.ul.pt/sectores/manual_anotacao.pdf).

The corpus samples come from several different sources:

- Spontaneous spoken corpus

- Informal dialogues and conversations collected for the project Português Fundamental (PF) (see description below), transcribed and published in Bacelar do Nascimento et al. (1987):

- Written corpus

- Fiction books - 70 titles of 53 Authors of the Portuguese Literature (XIX<sup>th</sup> and XX<sup>th</sup> centuries).

- Technical books - 39 titles of 38 Authors, published (end of the XX<sup>th</sup> century and XXI<sup>th</sup> century).

- Newspaper - several editions of year 2000 of the following newspapers: "A BOLA", "Diário de Notícias", "Expresso", "Jornal de Noticias" and "PUBLICO".

- Magazines - numbers 83 to 95 of the magazine "Revista do Instituto do Consumidor" (1999 and 2000).

- Varia - several articles from the "Enciclopédia Verbo", from scientific meetings proceedings, webpages, interviews published in the newspaper "O Primeiro de

1 Institutions that have been giving financial support to the CRPC: Fundação Calouste Gulbenkian, Junta Nacional de Investigação Científica e Tecnológica (JNICT) - Programma Estímulo em Ciências Sociais e Humanas, Fundação para a Ciência e Tecnologia (FCT) - Fundos Programáticos, Instituto Camões, União Latina, Caixa Geral de Depósitos, Comissão das Comunidades Europeias - LE-PAROLE Project. A net of public and private institutions is supplying data for CRPC (http://www.clul.ul.pt/english/frames.html).
Janeiro”, manuals for college students, final reports for bachelor training posts, etc.

### Spoken corpus

The Spoken corpus consists of informal conversations and is transcribed as follows:

- **Jornal** (newspaper) - 4,097,864 running words
- **Livrotel** (technical books) - 1,440,625 running words
- **Revista** (magazines) - 420,792 running words
- **Varia** (varia) - 812,599 running words
- **Jornal anotado** (tagged newspaper) - 336,151 running words
- **Livro anotado** (tagged books) - 125,434 running words
- **Revista anotado** (tagged magazines) - 25,908 running words
- **Varia anotado** (tagged varia) - 13,549 running words

**Total spoken corpus (ORAL)** - 105,964 running words

### Written corpus

The Written corpus is constituted as follows:

- **Jornal** (newspaper) - 4,097,864 running words
- **Livrotel** (technical books) - 1,792,590 running words
- **Revista** (magazines) - 420,792 running words
- **Varia** (varia) - 812,599 running words
- **Jornal anotado** (tagged newspaper) - 336,151 running words
- **Livro anotado** (tagged books) - 125,434 running words
- **Revista anotado** (tagged magazines) - 25,908 running words
- **Varia anotado** (tagged varia) - 13,549 running words

**Total written corpus (ESCRITO)** - 9,065,516 running words

**Total corpus (TOTAL)** - 9,171,480 running words

### Table 1: RL Corpus distribution and dimension

<table>
<thead>
<tr>
<th>Subcorpus</th>
<th>Subtotal (spoken) (ORAL)</th>
<th>Subtotal (written) (ESCRITO)</th>
<th>Total (TOTAL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jornal_RL (newspaper)</td>
<td>4,097,864</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Livrotel_RL (technical books)</td>
<td>1,792,590</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revista_RL (magazines)</td>
<td>420,792</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varia_RL (varia)</td>
<td>812,599</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jornal_anotado_RL (tagged newspaper)</td>
<td>336,151</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Livro_anotado_RL (tagged books)</td>
<td>125,434</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revista_anotado_RL (tagged magazines)</td>
<td>25,908</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Varia_anotado_RL (tagged varia)</td>
<td>13,549</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (spoken)</td>
<td><strong>105,964</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subtotal (written)</td>
<td></td>
<td><strong>9,065,516</strong></td>
<td>9,171,480</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td><strong>9,171,480</strong></td>
<td></td>
</tr>
</tbody>
</table>

This table shows the distribution and dimension of the RL Corpus into spoken and written subcorpora.

### ELAN CORPUS

The ELAN - European Language Activity Network (1998 - Program MLIS - 121) corpus was constituted within a LE-PAROLE project sequence and is composed of newspapers, techno-scientific books, periodicals and other (varia) texts samples, as follows in Table 2:

<table>
<thead>
<tr>
<th>Subcorpora</th>
<th>Subtotal (ELAN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jornal_ELAN (newspaper)</td>
<td>1,878,156</td>
</tr>
<tr>
<td>Livrotel_ELAN (techno-scientific book)</td>
<td>510,562</td>
</tr>
<tr>
<td>Revista_ELAN (periodical)</td>
<td>262,465</td>
</tr>
<tr>
<td>Varia_ELAN (other texts)</td>
<td>189,356</td>
</tr>
<tr>
<td>The entire corpus</td>
<td><strong>2,840,552</strong></td>
</tr>
</tbody>
</table>

### Table 2: ELAN Corpus distribution and dimension

This table shows the distribution and dimension of the ELAN Corpus into subcorpora.

### “PORTUGUÊS FUNDAMENTAL” PUBLISHED SAMPLE CORPUS

This Português Fundamental (PF) corpus is a spoken CRPC subcorpus, composed of 105,964 running words. It is constituted by the 140 published interviews recorded in the 70’s and representative of the whole PF corpus (Bacelar do Nascimento et al., 1987). It is available for download.

### ON-LINE QUERIES TO EUROPEAN PORTUGUESE LEXICONS

The European Portuguese has now a 26,443 lemma Frequency Lexicon with 140,315 different forms, with the minimum lemma frequency of 6, extracted from a relevant contemporary Portuguese corpus (16,210,438 running words). Each lemma is followed by morphosyntactic and quantitative information. The same information is given regarding each lemma token (inflected forms and some compounds). The lexicon indexations are listed in alphabetical order or decreasing frequency order, both available for on-line query or download at (http://www.clul.ul.pt/english/frames.html).

A more specific description is presented in the paper “Multifunctional Computational Lexicon Of Contemporary Portuguese: An Available Resource For Multitype Application” in this volume.

### OTHER AVAILABLE RESOURCES

C-ORAL-ROM - INTEGRATED REFERENCE CORPORA FOR SPOKEN ROMANCE LANGUAGES

(IST Programme - European Commission).

Considering that spoken language is not adequately represented in the currently existent LRs, specially when compared with the written language, the C-ORAL-ROM project intends to increase the LRs in that area, by establishing, building and making available a comparable corpus of spoken language of the four main romance languages, Spanish, Portuguese, French and Italian (with 300,000 words each language, covering both formal and informal speech). This multilingual corpus will soon be available in a multimedia edition on DVD, with text and
sound alignment, with standard tools for concordances extraction and acoustic signal analysis. Besides the quantitative and qualitative analysis and the several linguistic research studies that these materials provide for, it is worth to mention its usefulness in the creation of a representative multilingual resource designed for validation of HLT (Human Language Technologies).


The *Português Falado* corpus is constituted by authentic spoken documents and it aims mainly to develop the capacity of Portuguese language understanding and production between foreign students of medium or high-level Portuguese studies. The materials published in four CD-ROM's (sponsored by Instituto Camões) contribute to the observation and analysis of spoken Portuguese in its geographical varieties. Thus they are also very useful to teachers, translators, interpreters and researchers in general.

This spoken Portuguese corpus published in CD-ROM, with text-to-sound alignment, was collected among diverse speakers having Portuguese as mother language (mostly in Portugal and Brazil) or as second language (mostly in Africa and Asia). This corpus consists of informal conversations between acquaintances, friends or relatives as well as formal acts as, for instance, radio programs or conferences. In a total of 86 recordings, the texts exemplify the Portuguese spoken in Portugal (30), in Brazil (20), in the African countries with Portuguese as official language: Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe (5 each), in Macao (5), in Goa (3) and in East-Timor (3), corresponding to 8h44m of recordings and to 91.966 tokens (Figure 1). The recordings cover a period that goes from 1970 to 2001, and approximately 70% of them fall upon the last decade.

These samples of Portuguese varieties are distributed in the four CD-ROMs in the following way:

1 - Portugal (recordings from the nineties);
2 - Portugal (recordings from the seventies and the eighties), Macao, Sao Tome and Principe and East-Timor;
3 - Angola, Cape Verde, Guinea-Bissau and Mozambique;
4 - Brazil and Goa.

Finally, 94 speakers appear in the recordings; their characterizations (origin, sex, age, professional status, level of education) are visible on the header of each transcription, together with information about the place, date and situation in which the recording was made and other relevant type of information.

The orthographic transcriptions and the alignment between sound and the corresponding graphical representation were performed from the recordings. The materials are presented in a way that favors the use of self-learning processes. Aiming at the user to be able to hear the recording and to read simultaneously the respective transcription in the computer screen, a coloured light runs over the transcription of the sequence which is being listened. The user can control what he is listening, can repeat sequences or jump parts of the text (Gonçalves & Veloso, 2000). The 4 CD_ROM's are available at CLUL (fbacelar.nascimento@clul.ul.pt) or at the Instituto Camões (ded@instituto-camoes.pt).

It is still worth mentioning that, since this corpus was not collected having in mind a specific user profile, it will be useful not only for students and teachers but also for researchers, translators and interpreters, among others, which will be able to select and analyze the materials according to their own particular aims.

![Figure 1: Português Falado: corpus distribution and dimension](http://www.elda.fr)

**LE-PAROLE CORPUS**

The LE-PAROLE corpus is a 3 million word corpus, with the following constitution: newspapers (65%), books (20%), magazines (5%) and miscellaneous (10%). A 250,000 words subcorpus was morphosyntactically tagged and manually disambiguated. It is available, for sale, at ELRA catalogue - [http://www.elda.fr](http://www.elda.fr).
LE-PAROLE LEXICON
The PAROLE Lexicon counts 20,000 headwords morphosyntactically tagged and syntactically described. It is available, for sale, at ELRA catalogue - http://www.elda.fr.

SIMPLE LEXICON
3000 multilingual units from PAROLE Lexicon are semantically characterised (SIMPLE programme). It is available, for sale, at ELRA catalogue - http://www.elda.fr.

REFERENCES


