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[1]), a monitor corpus containing, at the present, more than 300 million words, taken by sampling from several types of written text (literary, newspaper, technical, didactic, juridical, parlamentary, 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** (Programme Lusitânia and Fundação Calouste Gulbenkian). The Sociedade Portuguesa de Autores (SPA), as partner of the project, had in charge copyright conditions. The corpus has 9.171.480 words, with a tagged subpart (See distribution in Table 1), and is available at:
http://www.clul.ul.pt/english/sectores/projecto_rld1.html.

- 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* sample corpus).

- **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 (XIXth e XXth centuries),
Technical books - 39 titles of 38 Authors, published (end of the XXth century and XXIst century),
Newspaper - several editions of year 2000 of the following newspapers: "A BOLA", "Diário de Notícias", "Expresso", "Jornal de Notícias" and "PÚBLICO",
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,

[1] Institutions that have been giving finantial support to the CRPC: Fundação Calouste Gulbenkian, Junta Nacional de Investigação Científica e Tecnológica (JNICT) - Programme Estímulo em Ciências Sociais e Humanas, Fundação para a Ciência e a 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).

interviews published in the newspaper "O Primeiro de Janeiro", manuals for college students, final reports for bachelor training posts, etc.

Spoken corpus transcribed and constituted by informal conversation:	105.964	
Subtotal (spoken) ORAL_RL		105.964
Written corpus constitution:		
jornal_RL (newspaper)	4.097.868	
livrolit_RL (fiction books)	1.792.590	
livrotec_RL (technical books)	1.440.625	
revista_RL (magazines)	420.792	
varia_RL (varia)	812.599	
jornal_anotado_RL (tagged newspaper)	336.151	
livro_anotado_RL (tagged books)	125.434	
revista_anotado_RL (tagged magazines)	25.908	
varia_anotado_RL (tagged varia)	13.549	
(tagged subcorpus) subcorpus_anotado_RL		501.042
Subtotal (written) ESCRITO_RL		9.065.516
TOTAL_RL		9.171.480

Table 1: RL Corpus distribution and dimension

ELAN CORPUS

The ELAN - European Language Activity Network (1998 - Programme MLIS) 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:

Subcorpora:	jornal_ELAN (newspaper)	1.878.156
	livrotec_ELAN (techno- scientific book)	510.562
	revista_ELAN (periodical)	262.465
	varia_ELAN (other texts)	189.356
The entire corpus	corpus_ELAN	2.840.552

Table 2: ELAN Corpus distribution and dimension

"PORTUGUÊS FUNDAMENTAL" PUBLISHED SAMPLE CORPUS

This *Português Fundamental* (PF) sample 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 is representative of the whole PF corpus (Bacelar do Nascimento et al., 1987). It is available for download.

ON-LINE QUERIES TO EUROPEAN PORTUGUESE LEXICON

MULTIFUNCTIONAL COMPUTATIONAL LEXICON OF CONTEMPORARY PORTUGUESE

(Funding Institution: JNICT / FCT – Programme PRAXIS XXI)

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

PORTUGUÊS FALADO - DOCUMENTOS AUTÊNTICOS: GRAVAÇÕES ÁUDIO COM TRANSCRIÇÃO ALINHADA (CD-ROM's)

(1995-1997 - Programme LINGUA/SOCRATES)

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) with text-to-sound alignement - contribute to the observation and analysis of spoken Portuguese in its geographical varieties. 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 its 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.

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.

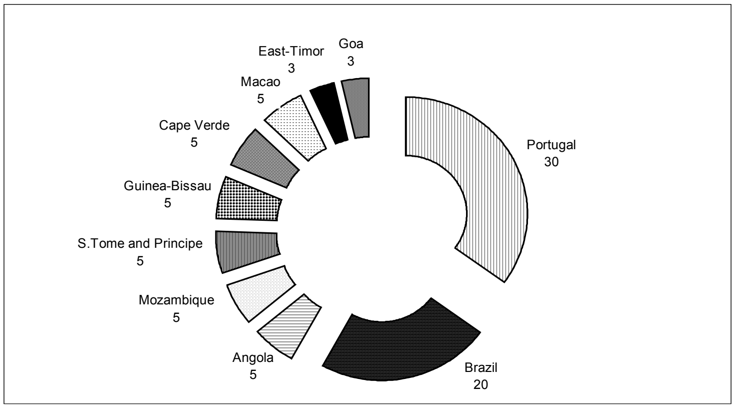

Figure 1: *Português Falado*: corpus distribution and dimension

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 four 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-to-sound alignment and prosodic segmentation, with standard tools for concordances extraction and acoustic signal analysis. Simultaneously, a book will be published with studies on specific phenomena of these spoken corpora. Besides the quantitative and qualitative analysis and the several linguistic research studies that the publication of these materials provide for, it is worth mentioning its usefulness in the creation of a representative multilingual resource designed for validation of HLT (Human Language Technologies).

LE-PAROLE CORPUS and LEXICON

(Programme Telematics Applications of Common Interest)

- 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.
- The LE-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

(Telematics Applications Programme)

3000 multilingual units from the LE-PAROLE Lexicon are semantically characterised. It is available, for sale, at ELRA catalogue - http://www.elda.fr.

REFERENCES

Bacelar do Nascimento, M. F. (2003). O lugar do *corpus* na investigação linguística. In A. Mendes et al. (Orgs.), Actas do XVIII Encontro da Associação Portuguesa de Linguística (pp. 601--605). Lisboa: Associação Portuguesa de Linguística e Edições Colibri.

Bacelar do Nascimento, M. F. (2003). O papel dos *corpora* especializados na criação de bases terminológicas. In I. Castro et al. (Eds.), Razões e Emoção, Miscelânea de Estudo em Homenagem a Maria Helena Mira Mateus, vol. 2 (pp. 167--179). Lisboa: Imprensa Nacional-Casa da Moeda.

Bacelar do Nascimento, M. F. (2002). Quelques considérations sur la constitution et l'exploitation d'un *corpus* de portugais parlé. In A. Scarano (a cura di) *Macro-syntaxe et pragmatique: l' analyse linguistique de l'orale* (pp. 295-302). Roma: Bulzoni Editore.

Bacelar do Nascimento, M. F. et al. (2002). The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus. In M. C. Rodrigues & C. S. Araujo (a cura di), Proceedings of the *Third International Conference on Language Resources and Evaluation* (pp. 2--10). Paris: ELRA.

Bacelar do Nascimento, M. F. (2001). Les études portugaises sur la langue parlée. In M. H. A. Carreira (Org.), Travaux et Documents, Les langues romanes en dialogue(s) (pp. 209--221). Vincennes Saint-Denis: Université Paris 8.

Bacelar do Nascimento, M. F. & Mota, M. A. (2001). Le Portugais dans ses variétés. In Revue Belge de Philologie et d'Histoire, 79, Fasc.3: Langues et Littératures Modernes (pp. 931-952). Bruxelles Société pour le Progrés des études philologiques et historiques.

Bacelar do Nascimento, F. (coord.) (2001). *Português Falado, Documentos Autênticos*, Gravações audio com transcrições alinhadas, em CD-ROM. Lisboa: Centro de Linguística da Universidade de Lisboa & Instituto Camões.

Bacelar do Nascimento, M. F. et al. (1987). Português Fundamental, vol. II - Métodos e Documentos, tomo 1 - Inquérito de Frequência. Lisboa: INIC, CLUL.

Bettencourt Gonçalves, J. (2000). Português Falado: variedades geográficas e sociais. In E. Gärtner et al. (Eds.), *Estudos de gramática portuguesa (1) (pp. 257-266).* Frankfurt am Main: TFM.

Bettencourt Gonçalves, J. & Veloso, R. (2000). Spoken Portuguese: Geographic and Social Varieties. In *Proceedings of the Second International Conference on Language Resources and Evaluation (*pp. 905-908). Athens, Greece: National technical University of Athens Press.

Callou, D. et al. (2003). A posição do adjectivo no sintagma nominal: duas perspectivas de análise. In S. F. Brandão & M. A. Mota (Orgs.), *Análise Contrastiva de Variedades do Português - Primeiros Estudos* (pp. 11—35). Rio de Janeiro: In-Fólio.

Mendes, A. et al. (2003). Reusing Available Resources for Tagging a Spoken Portuguese Corpus. In A. Branco et al. (Eds.), Tagging and Shallow Processing of Portuguese: worshop notes of TASHA'2003 (pp. 25--27). Lisboa: Departamento de Informática, Faculdade de Ciências da Universidade de Lisboa.

Pereira, L. A. S. & Bacelar do Nascimento, M. F. (2003). Contribuição para uma tipologia dos verbos portugueses frequentes em contexto: concordâncias do verbo contar. In *Como pôr os alunos a trabalhar? Experiências formativas na aula de Português* - 5º Encontro Nacional da APP (pp. 241-251). Lisboa.