

Portuguese corpora at CLUL

Maria Fernanda Bacelar do Nascimento, Luísa Pereira, João Saramago

Centro de Linguística da Universidade de Lisboa
Av. 5 de Outubro, N°85, 5°-6° 1050-050 LISBOA
{fbacelar.nascimento, luisa.alice.sp, j.saramago}@clul.ul.pt

Abstract

The *Corpus de Referência do Português Contemporâneo* (CRPC) is being developed in the Centro de Linguística da Universidade de Lisboa (CLUL) since 1988 under a perspective of research data enlargement, in the sense of concepts and hypothesis verification by rejecting the sole use of intuitive data. The intention of creating this open *corpus* is to establish an on-line representative sample collection of general usage contemporary Portuguese: a main *corpus* of great dimension as well as several specialized *corpora*. The CRPC has nowadays around 92 million words. Following the use in this area, the CRPC project intends to establish a linguistic database accessible to everyone interested in making theoretical and practical studies or applications.

The *Dialectal oral corpus of the Atlas Linguístico-Etnográfico de Portugal e da Galiza* (ALEPG) is constituted by approximately 3500 hours of speech collected by the CLUL Dialectal Studies Research Group and recorded in analogic audio tape. This *corpus* contains mainly directed speech: answers to a linguistic questionnaire essentially lexical, but also focusing on some phonetic and morpho-phonological phenomena. An important part of spontaneous speech enables other kind of studies such as syntactic, morphological or phonetic ones.

1. Corpus de Referência do Português Contemporâneo (CRPC)

The CRPC at the Centro de Linguística da Universidade de Lisboa is an electronically based linguistic *corpus* containing at the present 92 million words taken by sampling from several types of written speech (literary, newspaper, technical, scientific, didactic, economics, decisions of the supreme court of justice, parliament) and oral speech (formal and informal).

These samplings pertain to nacional and regional varieties of Portuguese, representing European, Brazilian, African, Macau, and East-Timor Portuguese. We intend to collect spoken Portuguese samples from some communities in India. From a chronological point of view our *corpus* contains texts from the second half of the XIX century up until now, mostly after 1970.

| |
|--|
| TOTAL DIMENSION 92 million words |
|--|

Material type

| |
|------------------------------|
| WRITTEN 90 333 172 |
| SPOKEN 1 767 163 |

Timespan

| |
|---------------------------------|
| BEFORE 1900 1 000 000 |
| 1901-1970 2 600 000 |
| AFTER 1970 88 500 335 |

Geographical distribution

| | |
|-------------------------------|--|
| PORTUGAL 85 308 811 | GUINEA-BISSAU 46 984 |
| ANGOLA 60 976 | MACAU 1 857 732 |
| BRAZIL 4 009 953 | MOZAMBIQUE 267 566 |
| CAPE VERDE 534013 | SÃO TOMÉ AND PRÍNCIPE 12 000 |
| EAST-TIMOR 2 300 | |

Source

| | |
|--|---|
| NEWSPAPER 55 098 899 | PARLIAMENT SPEECH 1 482 132 |
| BOOK 20 581 679 didactic - 1 968 143 literary - 8 841 143 tec-scient. - 9 772 063 | DECISIONS OF SUPREME COURT OF JUSTICE 1 809 400 |
| PERIODICAL 6 970 576 | LEAFLETS 322 538 |
| VARIA 3 904 756 | CORRESPONDENCE 163 192 |

Table 1: Composition and dimension of the CRPC

Corpus linguistics has become mainstream for researchers, teachers and students in the many areas of theoretical and descriptive linguistics, of language technology and of applied areas as translation (human and assisted), language teaching and learning, lexicography or information retrieval. It is recognized as an essential factor for the enhancement of communications and for facilitating access to information, in response to the basic needs of the Language Engineering in the framework of the multilingual Information Society.

Nowadays, through out the world, there is a growing interest in large *corpora* and *lexicon*¹ due to the extraordinary growth of affordable computer resources.

The running *corpus* CRPC is a resource and knowledge database made of authentic linguistic documents, organised in an electronic format already accessible to researchers, teachers, translators and to all, National and foreign, working on the Portuguese language to whom there is a need for reliable linguistic data.

These linguistic specific resources, closely related to each people's culture, in connection with important technologies for data and knowledge extraction, constitute an essential prerequisite for a large number of research projects and several types of development and applications, namely: new descriptions of the Portuguese language based on real data; contrastive studies between varieties of the Portuguese language aiming at finding factors of unity and diversity; contrastive studies between languages with similar linguistic resources; grammars; lexica and monolingual, bilingual or plurilingual dictionaries, as well as general and specialized dictionaries or conventional and electronic ones; terminologies; assisted translation materials; language teaching materials; developments and applications within language engineering, namely in what concerns processing systems, treatment and recognition of the natural language, language technologies and telecommunications.

The CRPC main goal is the continuous construction of a balanced *corpus* and its availability and dissemination to make this resource easily and friendly accessible. Besides the CRPC has already been used in numerous academic projects (mainly in MA and Ph.D studies) carried out in Portugal and abroad, and in other research projects, it is clear the major significance of such an aim.

In the very next times, a CRPC *subcorpus* with 3 million words will be soon distributable by ELRA Association and a CRPC *subcorpus* of other 3 million words will be available within ELAN project.

1.1. Applications

1.1.1. Lexicology and Lexicography

1.1.1.1. (1993/1996) Dicionário de Combinatórias do Português (DCP)

(Programme "Lusitania": Instituto Camões and Junta Nacional de Investigação Científica e Tecnológica (JNICT) - Project PLUS/C/LIN/816/93)

Taking a *subcorpus* of 12 millions words from CRPC, the extraction of all pairs of contiguous words and the pairs with one, two, three or four words between the two words of the pair was made. All these pairs are provided with the following information:

- Frequency and Distribution (oral, written, literature, newspaper, etc);
- Distance between the words of the pair;
- Mutual information;
- Localization of the *subcorpus*;

¹ Common examples are The Bank of English (English *corpus* with 200 million words), Frantext (French *corpus* with 240 million words), CREA (Spanish *corpus* with 200 million words by 1998).

- One line collocations of key-word.
- Groups of 3, 4, 5 words in fixed sequence.

1.1.1.2. (1997-) Léxico Multifuncional Computorizado do Português Contemporâneo

(Programme PRAXIS XXI, Project PRAXIS/2/2.1/759/95)

This Project aims to the construction of a general lexikon, of 30000 words, extracted from a *subcorpus* of CRPC, of 15 million words, of European Portuguese oral and written. Each lexical entry will contain quantitative information (Frequency and Distribution) and morphosyntactic classification.

1.1.2. Linguistic Engineering

In fonction of the existence of CRPC, the CLUL is, or was, partner of the following EC Commission Projects (DG-XIII).

1.1.2.1. (1993/94) Linguistic Research and Engineering Programms (LRE: MA), Project Network of European Reference Corpora (ML-85B-NERC2) Coordination: Consorzio Pisa Ricerche under the direction of Professor A. Zampolli.

The aim of this Project is to provide the European Commission with information about the future of linguistic *corpora* existing in Europa, having in view the harmonization of methodologies for selection, processing and analysis of the materials.

1.1.2.2. (1994/95) Multilingual Action Plan (LRE:MLAP), Project Preparatory Action for Linguistic Resources Organization for Language Engineering (PP-PAROLE: LRE-63-368-MLAP) Coordination: Consorzio Pisa Ricerche under the direction of Professor A. Zampolli.

The aim of this Project is to promote and prepare the establishment of substructures for creation, reuse and harmonization of resources within the domain of *corpora* and the creation of tools.

1.1.2.3. (1996/98) Telematics Applications of Common Interest, Project LE-PAROLE (LE-4017)

Coordination: Consorzio Pisa Ricerche under the direction of Professor A. Zampolli.

With this Project it is intended to provide the development of coherent and integrated model for reusing the *corpora* available in the countries of European Community, having in view the creation of lexica according a common model of linguistic description (flexion, morphosyntax, syntax), facilitating the multilingual contacts in the Community. It was build a 20 million words Portuguese *corpus* represented in SGML format from wich 250000 words have a morphosyntactic tagging.

1.1.2.4. (1998-) Semantic Information for Multifunctional Prurilingual Lexica (SIMPLE)

Coordination: Consorzio Pisa Ricerche under the direction of Professor A. Zampolli.

The goal of this project is to add semantic information to the set of harmonised multifunctional lexica built for 12 European languages by the PAROLE consortium. The

following types of semantic information, selected on the basis of the current state-of-the-art and the most urgent needs of LE systems, will be encoded for 10.000 semantic units: domain information, semantic class, template type, inheritance information, glossa, predicative representation, selectional restrictions, qualia structure, polysemous class.

1.1.2.5. (1998-) European Language Activity Network (ELAN) (MLIS Programme-121) Coordination: Professor C. Delcourt

The goal of this project is to make available a whole *corpus* representing the 30 European languages, with the dimension of 3 million words for each language. A Common *Corpus* Query Language will be developed, based on the CQL (*Corpus* Query Language) developed by partners of PAROLE and TELRI Associations and the creation of an user interface.

1.1.3. Learning/teaching Portuguese

1.1.3.1. (1995/1997) Projecto Português Falado, Variedades Geográficas e Sociais (EC Programme, (DG-XXII), LINGUA/SOCRATES - Contrat n° 94-09/1795/P-VB)

The objective of this Project is:

- the obtention, processing and analysis of samples of spoken Portuguese, spontaneous and formal, recorded in the last 25 years, including the varieties of Portugal, Brazil, Macau, East-Timor and Portuguese official language African countries, having in view to publish 83 texts in CD-ROM with sound and simultaneous orthographic transcription;
- the publishing of studies on spoken Portuguese (vocabulary, syntax, collocations lists,...).

1.2. Supporting Institutions

1.2.1. Institutions which have financed the CRPC

Fundação Calouste Gulbenkian, Fundação Oriente, Junta Nacional de Investigação Científica e Tecnológica (JNICT) – Programa Estímulo em Ciências Sociais e Humanas, Caixa Geral de Depósitos, Comissão das Comunidades Europeias – Projecto LE-PAROLE e União Latina

1.2.2. Institutions which provided materials for CRPC

Academia das Ciências de Lisboa; Agência Lusa; Assembleia da República; Caixa Geral de Depósitos; Centro de Informática do Ministério da Justiça; Coimbra Editora; DECO; Editora Colibri; Editora Nova Fronteira - Brasil; Editorial Verbo; Estação de Rádio TSF; Fundação Calouste Gulbenkian - Serviço de Bibliotecas e Apoio à Leitura; Instituto do Consumidor; Jornais portugueses: Expresso, O Público, Diário de Notícias, Diário Económico, Jornal de Notícias, A Bola, A Capital, O Independente, Jornal do Minho; Jornais de Cabo Verde: Correio Quinze, Novo Jornal, A Semana; Procuradoria-Geral da República; *Corpus* do Português Contemporâneo (Universidade Estadual Paulista - UNESP); Projecto NURC-BR (São Paulo e Rio de Janeiro); Projecto PEUL (Rio de Janeiro); Revistas: Grande Reportagem,

Ingenium, ProTeste, Máxima, Visão; Selecções do Reader's Digest; Sociedade Bíblica Portuguesa.

2. Dialectal oral corpus of the Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG)

At the present time the CLUL Dialectal Studies Research Group has collected approximately 3500 hours of recorded speech in 176 inquiries carried out in the Portuguese continental territory, 24 in the two insular archipels (7 in Madeira and 17 in Azores) and 10 in Spain (near the border). This *corpus* contains mainly directed speech: answers to a linguistic questionnaire specially built for a national Atlas. Thus it is essentially a lexical questionnaire. Nevertheless it also focuses on some phonetic and morpho-phonological phenomena. Since there is also an important part of spontaneous speech, the *corpus* enables other kind of studies, such as syntactic and morphological analysis or phonetic studies.

2.1. Applications

2.1.1. Atlas Linguístico-Etnográfico de Portugal e da Galiza (ALEPG)

The main application of the recorded dialectal oral *corpus* will be the publication of *Atlas Linguístico-Etnográfico de Portugal e da Galiza* (ALEPG) the project of the linguistic national atlas.

Meanwhile this *corpus* has been used for other international and national geolinguistic projects:

Atlas Linguarum Europae (ALE) (the linguistic atlas of all european languages)

Portugal participates in this project since 1975. Until now five volumes were published.

Atlas Linguistique Roman (AliR) (the linguistic atlas of the romance domain)

This project had appeared in the late eighties and its target is to provide a wide-spread vision of the linguistic situation of the romance languages in Europe. In 1996 the first volume was published. The second is for publication.

Atlas Linguístico e Etnográfico dos Açores (ALEAç) (the linguistic atlas of the azorean archipel)

This medium-term project emerged within the context of ALEPG. Its aim is the publication of the material collected in the 17 points of the archipel which constitutes the inquiry net. During 2000 the first volume will be published.

Atlas Linguístico do Litoral Português (ALLP) (Linguistic Atlas of the Portuguese Coast).

In the oral *corpus* there is recorded material for a very specific project that aims at studying the lexicon related to fishing daily life along the portuguese seaboard. The dialectal questionnaire has about 1200 questions on the following semantic fields: fishing and fishing procedures; boats and sailing; crew and fishing trade; sea fauna and flora; seaboard and sea; weather phenomena. Forty localities set up the inquiry net: 23 on the continental coast, 5 in Madeira archipel and 12 in Azores archipel.

A pre-print of 200 lexical maps on sea fauna and flora (with an alphabetical list of contents) is achieved for the continental domain. It includes the collected data concerning the different species and the general features

of fauna, with data phonetic transcriptions. Each map deals with just one concept or one biological species and includes a set of comments on specific features related to the concept or the lexical information. In addition, species not regularly researched are also considered in the maps comments.

2.1.2. Other applications

The parts of spontaneous speech existing in the *corpus* enables other kind of studies such as syntactic, morphological or phonetic ones.

For the time being two projects are studying the recorded material of this point of view:

2.1.2.1. *Corpus Dialectal com Anotação Sintática (CORDIAL-SIN)* (*Corpus* of Portuguese Dialects Syntactically Annotated)

This project is aimed at developing and enhancing research activity on syntactic dialect variation.

The project outcome will be an electronic medium-size database – of about 30.000 sentences – integrating a detailed syntactic annotation.

Studies on some aspects of the syntax of Portuguese dialects will be carried out, in view of publication, by the research team – topic, focus and wh-constructions, clitic placement, subordinate adverbial clauses.

2.1.2.2. *Variantes Inflexionais do Verbo no Português Continental Falado* (Inflexional Variants of the Verb, in Spoken Continental Portuguese)

This project intends to make the inventory of the variants of verbal inflexion observed in the continent in order to establish the variant patterns analysing and characterising them mostly in terms of morphological and phonological features and to define the main dialectal/geographical areas of each inflexion pattern and the inflexion *continuum* between areas.

On one hand is also previewed to compare the patterns of the North of Portugal with those of the gallician territory; on the other hand with those of Mozambican portuguese and brazilian portuguese.

lexicografia, terminologia, *ALFA, Revista de Linguística*, S. Paulo: UNESP, 42:183-203.

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. e L.A.S. Pereira 1996. Dicionário de Combinatórias do Português: associações frequentes observadas num *corpus* de Português contemporâneo. In I.H. Faria e M. Correia (orgs.), *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística, Dicionários*, Lisboa: APL, II:43-54.

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.

Carreira, M.H.A. 1997. *Modalisation Linguistique en situation d'interlocution : proxémique verbale et modalités en portugais*, Leuven-Paris, Peeters.

Cruz, M.L.S. e J. Saramago 1999. Açores e Madeira: autonomia e coesão dialectais. In I.H. Faria (org.), *Lindley Cintra, Homenagem ao Homem, ao Mestre e ao Cidadão*:707-738.

Segura da Cruz, M.L. 1996. Os *corpora* dialectais do CLUL : sua caracterização e objectivos. In Bacelar do Nascimento, M.F. et alii (orgs.), 151-158.

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.

3. 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 sintáticas 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. 1998. Resultados do Projecto Dicionário de Combinatórias do Português, O estado da arte nas ciências do léxico: lexicologia,