The Reference Corpus of Contemporary Portuguese and Related Resources

Maria Fernanda Bacelar do Nascimento, Amália Mendes, Sandra Antunes, Luísa Pereira

[A] 1. Introduction

The extraordinary growth of computer applications, particularly over the last two decades, has enabled the easy compilation and exploration of large corpora and lexica. These linguistic resources play a fundamental role in the areas of theoretical linguistics and natural language engineering. Combining these two areas of knowledge can, in fact, result in the development of a large number of applications, such as new and straightforward descriptions of languages based on real data; contrastive studies between varieties of a particular language aiming at finding factors of unity and diversity; cross-linguistic contrastive studies; grammars; lexica and dictionaries; terminologies; assisted translation materials; language teaching materials; computer tools and applications for processing natural language.

Having this principle in mind and following the tradition at the Centre of Linguistics of the University of Lisbon (CLUL)[i] of collecting and studying real language data, a large electronic corpus – the Corpus de Referência do Português Contemporâneo (Reference Corpus of Contemporary Portuguese, CRPC) – is being compiled at CLUL since 1988. The CRPC currently contains approximately 310 million words, searchable through a user-friendly interface, and it is envisaged as a monitor corpus (from which one can extract balanced subcorpora) that can serve as a sample of the Portuguese language (both in its written and spoken varieties).

In the next sections, we will describe the CRPC and how it forms the basis for important resources developed at CLUL.

[A] 2. The Reference Corpus of Contemporary Portuguese

The CRPC is a large electronic Portuguese corpus that has been under development at CLUL since 1988[ii]. Presently, this corpus contains 311.4 million words from written and spoken language, reflecting both regional and national varieties of Portuguese (Portugal, Brazil, Angola, Cape Verde, Guinea-Bissau, Mozambique, Sao Tome and Principe, Goa, Macao and East Timor), distributed as shown in Figure 1.

[Insert Figure 1(Figure 01_Antunes_et_al) here]

Figure 1: Varieties covered by the CRPC and number of tokens for each

The CRPC covers written texts (309.8 million words) from different text types (e.g., newspapers, books, periodicals, parliament sessions, decisions of the Supreme Court of Justice, leaflets, correspondence, miscellaneous) and spoken transcriptions (1.6 million) of informal and formal sessions (the latter of which includes radio and television programmes – features, news, talk shows, interviews, sport – political speeches and debates, conferences, preaching and teaching). From a chronological point of view, the CRPC includes texts from the second half of the 19th century up until now, most of which were produced after 1970 (Bacelar do Nascimento et al., 2000; Bacelar do Nascimento, 2000). Table 1 shows the constitution of the CRPC and its broad variety of text types.

[Insert Table 01 (Table 01_Antunes_et_al) here]

Table 1: Composition of the CRPC

In order to enhance access to our corpus, the written part of the CRPC has recently been enriched with linguistic information and metadata, and undergone a full process of cleaning and preparation for online queries. CRPC is now available on the CQPWeb platform (Mendes et al., 2012; Généreux et al., 2012)[iii], which enables extensive search options. Amongst the available options are:

(i) restricted queries over specific text types or text varieties;

(ii) queries for words, regular expressions, lemmas, part-of-speech tags, nominal chunks;

(iii) sort and download of concordances results;

(iv) frequency and information on the distribution of the search item in the texts of the corpus;

(v) collocations information;

(vi) lexical comparison of subcorpora;

(vii) subcorpora customization.[iv]

As an example, Figures 2 and 3 below show a query for the common noun lemma poder (power), restricted by both variety (Portugal only) and register (newspaper), with a concordance display.

[Insert Figure 2 (Figure 02_Antunes_et_al) here]

Figure 2: Search for the common noun lemma poder (power) in the CRPC web interface

[Insert Figure 3 (Figure 03_Antunes_et_al) here]

Figure 3: Concordances for the common noun lemma poder (power) from the CRPC web interface

The spoken subpart of the CRPC has been developed under specific projects and is constituted by several subcorpora that will be described in section 3.1. Almost all of them are publicly available, either freely or for purchase.

[A] 3. Related Resources

The CRPC has been used in numerous national and international research projects and academic studies, many of which resulted in linguistic resources that are available online[v]. Some of these resources deserve particular attention and will be described in the following sections.

[B] 3.1. Spoken Corpora of European Portuguese

[C] 3.1.1. Português Fundamental (1984)

Having Français Fondamental as a reference (as well as its Spanish counterpart), Português Fundamental was the first spoken corpus collected at CLUL, between 1970 and 1974. At the time, the task of establishing the vocabulary essential to communication when teaching a foreign language was performed by teachers and textbooks authors guided only by their intuition. In order to change this method, the project aimed to provide information on the Portuguese vocabulary often used in everyday life situations. To collect this vocabulary, two corpora were compiled: the Frequency Corpus (Corpus de Frequência) and the Availability Corpus (Corpus de Disponibilidade).

(i) The Frequency Corpus (Corpus de Frequência)

This heterogeneous spoken corpus aimed to be representative of the Portuguese language, and, as in all of these types of corpora, particular attention was given to some sociolinguistic characteristics of the informants. They were intended to be from all districts of continental Portugal and islands, of different ages and diversified social and professional backgrounds.

From a total of 1,800 recordings (approximately 500 hours) of spontaneous spoken communication, on different themes of everyday life, a total of 1,400 were selected and 700,000 words were transcribed, making up the Frequency Corpus. From this corpus, the total list of occurring word forms was extracted (25,107 different word forms), together with their frequency. This list was lemmatized and used to define the set of lemmas with frequency equal to or greater than 40 (1,179 lemmas), thereby forming the Frequency Vocabulary.

[Insert Table 02 (Table 02_Antunes_et_al) here]

Table 2: Example of word lists from the Fundamental Vocabulary, sorted alphabetically by lemma (including contracted forms) and by reverse frequency

The sample of the transcriptions of the spoken corpus published in 1987 (140 recordings) is freely available for download[vi]. The transcriptions of this sample have recently been revised according to the CHAT guidelines[vii], text-to-sound aligned with the EXMARaLDA software (Schmidt, 2012), and automatically lemmatized and annotated with PoS information. This updated version of the Português Fundamental Corpus is freely available for research purposes on the ELRA catalogue[viii].

(ii) The Availability Corpus (Corpus de Disponibilidade)

Although special care was taken to cover a range of different themes in the Frequency Corpus, some vocabulary had nonetheless a very low or even zero occurrence in the recordings, due to the fact that these lexical items were only used in specific contexts, and also because they were often replaced by deictics or other elements in the conversation. These cases were then addressed through a selection of 30 topics with lower probability of occurrence in spontaneous spoken discourse but admittedly essential to communication (e.g., politics, working relationships, the human body, health and illness, professions and trades). These topics were addressed in a supplementary survey, called the Availability Corpus (Corpus de Disponibilidade), which was then implemented between 1970 and 1974. It consisted of questionnaires where informants were asked to identify what they felt were the most significant nouns, adjectives and verbs related to the topics at hand. The result is a set of answers to questionnaires, called the Availability Corpus (which is not, in fact, a textual corpus). The questionnaires further lead to a vocabulary of 481,800 words from specific topics, named the 'Availability Vocabulary'.

After 1974, a supplementary survey was made, with the purpose of covering themes that were considered sensitive before the revolution of April 25[ix] because of the political censorship.

[C] 3.1.2. C-ORAL-ROM – Integrated Reference Corpora for Spoken Romance Languages (2004)

The C-ORAL-ROM[x] corpus was developed to answer the need to increase the spoken language resources for Romance languages[xi]. A multilingual corpus of spontaneous spoken language of the four main Romance languages (French, Italian, European Portuguese and Spanish) was compiled and made available. The corpus comprises 1,200,000 words (300,000 words for each language), covering both formal and informal speech. A similar approach was recently followed for Brazilian Portuguese and resulted in the C-ORAL-BRASIL, as described by Raso and Mello in this volume.

The European Portuguese corpus contains 153 recordings, lasting 30 hours in total. The corpus design (with a matrix for all the languages) is represented in Table 3.

[Insert Table 03 (Table 03_Antunes_et_al) here]

Table 3: The C-ORAL ROM European Portuguese corpus contents

This resource for spoken Romance languages is a benchmark for corpus design, dialogue representation, prosodic annotation, part-of-speech (PoS) tagging, multimedia storage and speech analysis. It comprises several components:

(i) a multimedia corpus, containing, for each text: (a) the acoustic source; (b) the orthographic transcription, in CHAT format and enriched with the tagging of terminal and non-terminal prosodic breaks; (c) session metadata containing essential information for speakers, recording situation and session; (d) text-to-sound synchronization, i.e. the alignment between the acoustic source and the transcribed utterances; (e) a second orthographic transcription with lemma and PoS tags of each form in the transcribed texts; (f) frequency lists for both forms and lemmas;

(ii) software tool for speech analysis, with simultaneous access to acoustic and textual information (WinPitch Corpus, developed by Pitch France[xii]);

(iii) a concordancer tool, which allows searches within both text-only and PoS-tagged files (Contextes, developed by Jean Véronis[xiii]);

(iv) appendixes containing descriptions of the four subcorpora as well as the procedures followed and choices made by each team during corpus design and preparation (e.g., orthographic transcription, lemmatization, tagging), in addition to comparative linguistic studies on lexical and structural strategies in the four languages, and models and standard linguistic measures of spoken language variability.

C-ORAL-ROM is available in two versions: (i) one with permission to explore the materials, as described above, available on eight DVDs, distributed by ELRA[xiv]; (ii) an encrypted version (that does not allow for the full extraction of concordances, for example), available on one DVD that accompanies the 2005 book published by John Benjamins (Cresti and Moneglia, 2005; Bacelar do Nascimento et al., 2005)[xv].

[B] 3.2 Corpora of Portuguese Varieties

[C]3.2.1 Spoken Portuguese – Geographical and Social Varieties (2001)

Considering that the use of authentic spoken texts in the teaching of Portuguese as a foreign language was not a common practice (written texts were often used to reproduce spontaneous speech), the main goal of the Spoken Portuguese corpus[xvi] was to collect and transcribe recordings of the different varieties of the Portuguese language in the world, thereby contributing to the improvement of the understanding and production skills in students of Portuguese as a second language.

This corpus represents real communication by sociolinguistically diverse speakers having Portuguese as their mother tongue or as a second language. It consists of informal conversations between acquaintances, friends and relatives as well as formal acts (radio programmes or conferences), in a total of 86 recordings (8h44m) and 91,966 tokens.

The corpus covers the Portuguese spoken in Portugal (30 transcribed recordings), Brazil (20), Angola (5), Cape Verde (5), Guinea-Bissau (5), Mozambique (5), Sao Tome and Principe (5), Macao (5), Goa (3) and East Timor (3), covering a period that goes from 1970 to 2001, 70% of which were produced between 1990 and 2001.

Users can explore this corpus to improve their listening skills, particularly their ability to understand different varieties of Portuguese not limited to spoken aspects (e.g., pronunciation, prosody, intonation contour, accent) but also including morphological, lexical, syntactic and discursive characteristics (Bettencourt Gonçalves and Veloso, 2000).

This resource was first published on CD-ROMs that included the recordings, the orthographic transcriptions, text-to-sound synchronization and metadata information about the speakers (e.g., origin, sex, age, professional status, level of education) as well as the place, date and situation in which the recording was made (Bacelar do Nascimento, 2001a). All these materials are also freely available for download at the project webpage[xvii].

The transcriptions of the Spoken Portuguese corpus have also been revised according to the CHAT guidelines, text-to-sound aligned with EXMARaLDA software, and automatically lemmatized and annotated with PoS information. This new version is freely available for research purposes on the ELRA catalogue[xviii].

[C] 3.2.2 Africa Corpus

Given the notorious lack of studies on African varieties of Portuguese, two interrelated projects[xix] were designed and conducted to fill this gap, namely Linguistic Resources for the Study of African Varieties of Portuguese (2006) and Properties of African Portuguese Varieties compared with European Portuguese (2008). Both provide resources for a description of five varieties, simultaneously contributing to a better understanding of the differences in spoken and written productions of native speakers from several countries. Specifically, the main goal of the Linguistic Resources for the Study of the African Varieties of Portuguese project was to constitute, analyze, and make available online a 3.2-million-word corpus of written and spoken texts, including five different subcorpora comparable in size, time frame, and genre, 640,000 words each, corresponding to the varieties of Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe, as shown in Tables 4 and 5.

[Insert Table 04 (Table 04_Antunes_et_al) here]

Table 4: Corpus design of African Portuguese varieties, by country

[Insert Table 05 (Table 05_Antunes_et_al) here]

Table 5: Design of the African Portuguese varieties corpora, written and spoken registers

These corpora were automatically annotated for lemma and PoS, and some difficult cases prone to automatic error tagging were manually revised. The Africa Corpus allows for the study of each African variety, but also for inter-corpora comparative studies of the varieties, which point, on the one hand, to the existence of a common core vocabulary and grammar across the varieties, with variations that result from discursive and pragmatic particularities, and on the other hand, to aspects of linguistic unity or diversity that characterize the spoken Portuguese of all five African countries (Bacelar do Nascimento et al., 2008a; Bacelar do Nascimento et al., 2008b). The analysis of these comparable subcorpora have yielded the following results:

(i) lists of lemmas and forms with frequency data, broken down by subcorpus and by genre (Table 6);

(ii) contrastive word indexes (lemmas and forms) that occur in each subcorpus with frequency data, divided by genre (Table 7);

(iii) comparative description of the vocabulary of the subcorpora – word formation processes and syntactic and morphosyntactic phenomena – as a result of quantitative and statistical analysis;

(iv) comparative study between the linguistic processes of each variety and those of European Portuguese (Table 8).

These lists and comparisons are freely available at the project webpage[xx].

[Insert Table 06 (Table 06_Antunes_et_al) here]

Table 6: Example of word indexes (lemma and form) that occur in Angola

[Insert Table 07 (Table 07_Antunes_et_al) here]

Table 7: Frequency of some forms in the five subcorpora in spoken and written registers

[Insert Table 08 (Table 08_Antunes_et_al) here]

Table 8: Examples of a comparative analysis of each variety with European Portuguese

[B] 3.3 Manually annotated subcorpora

[C] 3.3.1. LE-PAROLE Corpus (1998)

The LE-PAROLE corpus is the result of an European initiative to develop corpora and lexica for all European Union (EU) languages according to mutual design and composition principles, using both linguistic and computer resources already available in the EU countries. The use of common tools and integrated models assured multilingual resources and comparable studies (for more details, see the project webpage[xxi]).

The languages involved in the LE-PAROLE project are Belgian, French, Catalan, Danish, Dutch, English, French, Finnish, German, Greek, Irish, Italian, Norwegian, Portuguese and Swedish. For each language, a 3-million-word corpus was built on a shared set of design, markup and annotation, including a 250,000-word corpus tagged with PoS annotation (Bacelar do Nascimento et al., 1998).

The LE-PAROLE corpora were classified and encoded according to the common core Parole Encoding Standard. There were agreed parameters for time of production (no texts older than 1970 were allowed), as well as for publication medium (inclusion of

specific proportions of texts from a closed set of categories: 'Book', 'Newspaper', 'Periodical' and 'Miscellaneous'). The content of the Portuguese corpus in terms of percentage of tokens is as follows:

(i) Newspaper: about 65%, from 1996 to 1997, 3 publications;

(ii) Book: about 20%, 12 titles from 3 publishing houses;

(iii) Periodical: about 5%, 7 weekly issues of 1 title, from 1996; and

(iv) Miscellaneous: about 10%, 8 titles.

As for the corpus annotation, an equal proportion of the corpus (up to 250,000 running words) was PoS-tagged for each language, based on a common core PAROLE tagset that was extended to include a set of language-specific features. Disambiguation was checked manually. The Portuguese corpus (raw and annotated) is available for sale on ELRA's catalogue[xxii].

For most of the languages involved, a lexicon of 20,000 entries was also implemented (see section 3.4.2).

[C] 3.3.2. CINTIL Corpus (2006)

CINTIL – Corpus Internacional do Português[xxiii] – is a corpus of Portuguese made up of 1 million annotated tokens, hand-checked by human expert annotators (Branco and Silva, 2004; Barreto et al., 2006). The corpus is the result of a joint enterprise between CLUL and FCUL (Faculty of Sciences of the University of Lisbon) in the scope of the TagShare project[xxiv]. The annotation comprises information on part-of-speech, open class lemma and inflection, multiword expressions (adverbial expressions and closed PoS classes), multiword proper names (for named entity recognition) and specific tags for spoken discourse (to account for extralinguistic and paralinguistic elements and

fragmented words)[xxv]. The corpus contains both written (58%) and spoken (42%) texts, as presented in Table 9.

[Insert Table 9 (Table 9_Antunes_et_al) here]

Table 9: Design of the CINTIL corpus

The corpus is available online for concordancing through a user-friendly interface that allows for both simple and advanced search options. Amongst the available search options are: simple orthographic forms; regular expressions; PoS; nominal and verbal inflection; named-entities; metadata. Furthermore, the complete corpus is available for sale on ELRA's catalogue[xxvi].

[B] 3.4 Lexica

[C] 3.4.1 Fundamental Vocabulary of Portuguese

As mentioned in section 3.1.1, the Português Fundamental corpus was used to identify the fundamental lexicon for European Portuguese. The Frequency Corpus (spoken corpus) and the Availability Corpus (questionnaires) were the source of two specific vocabularies, which resulted in the Fundamental Vocabulary of Portuguese, with 2,217 words, published in 1984 (Bacelar do Nascimento et al., 1984). Two further volumes were published containing a detailed description of the methods used in compiling, analyzing and establishing this vocabulary, and also a set of documents resulting from these collections and analysis which included the following: a sampling of the transcriptions of the recorded conversations; lemmatized lists with frequency information, sorted alphabetically and by decreasing frequency, extracted from both

corpora; and a joint list of the lemmas of these two corpora (Bacelar do Nascimento et al., 1987a, 1987b).

[C] 3.4.2 PAROLE/SIMPLE Lexicon (2000)

The PAROLE lexica cover 12 of the 15 languages included in the LE-PAROLE project: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Portuguese, Spanish and Swedish. The lexica contain 20,000 entries each per language, including both PoS and syntactic information, according to the PAROLE common encoding standards. The Portuguese lexicon (Marrafa et al, 1999) is available for sale on ELRA's catalogue[xxvii].

A follow-up of this work was undertaken by the SIMPLE (Semantic Information for Multifunctional Plurilingual Lexica) project[xxviii], which aimed to add semantic information to a set of morphosyntactic units of the PAROLE lexica. This resulted in a set of mutual multifunctional syntactic lexica. The attributes of the semantic units include examples, definitions, semantic features (domain, inheritance template, partial qualia information) and relations (synonymy, hyponymy, and predicative representation). The resulting SIMPLE Portuguese lexicon contains 10,438 such units[xxix]. The Portuguese PAROLE lexicon is also available on the ELRA catalogue[xxx].

[C] 3.4.3. Multifunctional Computational Lexicon of Contemporary Portuguese (2000)

Another lexical resource based on a subcorpus of the CRPC is the Multifunctional Computational Lexicon, which provides frequency, lemma and PoS information, and followed what were at the time state-of-the-art standards (cf. Halliday, 1993, p. 1 apud

Bacelar do Nascimento, 2001b). Special care was taken with the design of the subcorpus that gave rise to the lexicon: the corpus contains 16,210,438 words, from both spoken (856,195 words) and written (15,354,243 words) registers, as shown in Figure 4. From this subcorpus, 26,443 lemmas were then extracted, amounting to 140,315 tokens (with a minimum lemma frequency of 6), which constitute the lexicon.

[Insert Figure 04 (Figure 04_Antunes_et_al) here]

Figure 4: Corpus design for the extraction of the Multifunctional Computational Lexicon of Contemporary Portuguese

All word forms were automatically tagged and lemmatized and manually revised (Bacelar do Nascimento, 2001b; Amaro and Barreto, 2004). Thus, each token and lemma is followed by both morphosyntactic and quantitative information regarding the number of occurrence in the corpus. This quantitative information is presented both as a probability value (to account for the fact that there is an error rate regarding the PoS tagging), and also as exact values of frequency. The lexicon entries are also listed in both alphabetical and reverse frequency order, as shown in Table 10.

[Insert Table 10 (Table 10_Antunes_et_al) here]

Table10: Example of the two lists of the Multifunctional Computational Lexicon of Contemporary Portuguese sorted alphabetically by lemma and by decreasing frequency

All the data is available for download at the project webpage, where more detailed information on the project can also be found[xxxi].

[C] 3.4.4 COMBINA-PT - Word Combinations in Portuguese Language (2006)

Another lexical resource based on the CRPC is a lexicon of multiword expressions (Mendes et al., 2006), developed in the scope of the project COMBINA-PT[xxxii]. As it is widely known, the lexicon does not consist mainly of single lexical items but appears to be populated with numerous chunks, more or less predictable, though not fixed (Firth, 1955). In fact, the availability of large amounts of textual data and the development of computer technologies and corpus-based approaches has enabled the identification of complex patterns of word associations, showingthat the speakers use a large number of prefabricated phrases that constitute single choices (Sinclair, 1991). For the extraction of lexical associations, a subcorpus of the CRPC was designed, with 50-million written words from different genres, as illustrated in Table 11.

[Insert Table 11(Table 11_Antunes_et_al) here]

Table 11: Corpus design for the extraction of lexical associations

The extraction of significant associations was implemented through a software application that extracts groups of 2, 3, 4 and 5 tokens from the corpus, together with several types of information, including the following: (i) the number of elements of the group; (ii) the distance between the elements of the group (groups of 2 tokens can be either contiguous or be separated by a maximum of 3 tokens (e.g., conjuntura internacional (international conjuncture); conjuntura económica internacional (international economic conjuncture)), whereas groups of more than 2 tokens must be contiguous); (iii) the frequency of the group at a specific distance and in all occurring distances; (iv) the frequency of each element of the group in the corpus; (v) lexical association measure (the groups automatically extracted are sorted using the lexical association measure Mutual Information (MI)[xxxiii]); (vi) concordance lines (in KIWIC

format) of the group in the corpus, together with an index code pointing to the occurring position in the corpus.

A sample of the large candidate list extracted from the corpus was hand-checked. This sample was established using the best MI values, ranging between 8 and 10 (for details on the results of the MI statistical measure, see Evert and Krenn (2001) and Pereira and Mendes (2002)). Throughout manual validation, multiword expressions that presented some syntactic and semantic cohesion were selected, paying particular attention to four important aspects:

i) lexical and syntactic fixedness that can be observed through the possibility of replacing elements, inserting modifiers, changing the syntagmatic structure or gender/number features;
ii) total or partial loss of compositional meaning, which means that the meaning of the expressions can not be predicted by the meaning of the parts;
iii) frequency of occurrence, which may reveal sets of favoured co-occurring forms, i.e., expressions that may be semantically compositional but occur with higher frequency than any other alternative expression of the same concept, which could point to an initial stage of lexicalization;
iv) syntactic category of the groups (verbal phrases, noun phrases, sentences).

The lexicon covers multiword expressions with different degrees of lexicalization, ranging from idiomatic expressions (i.e., fully lexicalized) with lexical and syntactic restrictions (deitar cedo e cedo erguer dá saúde e faz crescer (early to bed and early to rise makes a man healthy, wealthy and wise); a sangue frio (in cold blood)) to collocations, i.e., non idiomatic expressions where the elements reveal a tendency to co-

occur in certain contexts (lufada de ar fresco (breath of fresh air); condenar ao fracasso (doomed to failure)).

The lexicon of multiword expressions[xxxiv] is organized as follows: each multiword expression in the lexicon is linked to the lemma of its node word, i.e., a single word from different PoS categories (e.g. fogo (fire)) and is also linked to a group lemma, which corresponds to the canonical form of the expression (e.g. arma de fogo (fire weapon)) and covers all the variants of the multiword expressions that occurred in the corpus (e.g. arma de fogo (fire weapon); armas de fogo (fire weapons)). In all, the lexicon comprises 1.180 main lemmas, 14.153 group lemmas and 48.154 word combinations (Bacelar do Nascimento et al., 2006; Mendes et al., 2006).

This resource may be useful in several areas, including psycholinguistics (development of hypothesis about the representation of an individual's mental lexicon, semantic memory and cognitive processes in general), lexicography (improvement of coverage in modern dictionaries), second-language acquisition (enhancement of acquiring the Portuguese significant word combinations, which will make the student's speech and writing sound more natural), and computational linguistics (development and evaluation of language processing tools capable of dealing with expression-specific issues, like automatic unit recognition or tagging and parsing problems).

[A] 4. Conclusion

Over more than twenty years, the CRPC (Corpus de Referência do Português Contemporâneo) has been enlarged and updated for new technologies, having become

the reference monitor corpus of Portuguese that it is today. It has been widely used both in linguistic research projects and in the development of software tools for the computational processing of Portuguese. The most relevant of these projects were mentioned above, but many more were carried out at CLUL, along with numerous academic studies that also used the CRPC as the basis for their research.

To continue improving this large corpus, future work includes: (i) adding more searchable metadata tags on the CQPWeb platform; (ii) introducing a language spotter for the few pockets of foreign language present in the corpus; (iii) enlarging linguistic annotation to cover nominal and verbal inflection and addressing issues related to multiword expressions; (iv) revising the design of the corpus to improve representativeness; and (v) contacting publishers and authors for copyright clearance for making a part of the CRPC freely available.

[A]5. References

Amaro, R. and Barreto, F. (2004), 'Multifunctional Computational Lexicon of Contemporary Portuguese: an available resource for multitype applications', Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, pp. 1075-1078.

Bacelar do Nascimento, M. F., Garcia Marques, M.L. and Segura da Cruz, M. L. (1984), Português Fundamental. Vocabulário e Gramática, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. F., Garcia Marques, M. L. and Segura da Cruz, M. L. (1987a), Português Fundamental. Métodos e Documentos. Inquérito de Frequência, Vol. 1, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. F., Garcia Marques, M. L. and Segura da Cruz, M. L. (1987b), Português Fundamental. Métodos e Documentos. Inquérito de Disponibilidade, Vol. 2, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. F., Marrafa, P; Pereira, L. A. S., Ribeiro, R., Veloso, R. and Wittmann, L. (1998), 'LE-PAROLE – Do corpus à modelização da informação lexical num sistema-multifunção', Proceedings of XIII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa, pp. 115-134.

Bacelar do Nascimento, M. F. (2000), 'O Corpus de Referência do Português Contemporâneo e os projectos de investigação do Centro de Linguística da Universidade de Lisboa sobre variedades do português falado e escrito', in Gärtner, E., Hundt, C. and Schönberger, A. (eds.), Estudos de Gramática Portuguesa (I), Biblioteca Luso-Brasileira, Centro do Livro e do Disco de Língua Portuguesa, Frankfurt am Main, pp. 185-200.

Bacelar do Nascimento, M. F., Pereira, L. and Saramago, J. (2000), 'Portuguese corpora at CLUL', Proceedings of the Second International Conference on Language Resources and Evaluation (LREC), Athens, Greece, Vol. III, pp. 1603-1607.

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

Bacelar do Nascimento, M. F. (2001b), 'Um novo léxico de frequências do português', in Volume de Homenagem ao Professor Herculano de Carvalho.

Bacelar do Nascimento, M. F., Bettencourt Gonçalves, J., Veloso, R., Antunes, S., Barreto, F. and Amaro, R. (2005), 'The Portuguese corpus', in E. Cresti and M. Moneglia (eds.), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, John Benjamins Publishing Company, Studies in Corpus Linguistics nº 15, Amsterdam/Philadelphia, pp. 163-207 (with DVD).

Bacelar do Nascimento, M. F., Mendes, A. and Antunes, S. (2006), 'Typologies of multiword expressions revisited: A corpus-driven approach' in Y. Kawaguchi, S. ZAIMA and T. Takagaki (eds.), Spoken Language Corpus and Linguistic Informatics, Jonh Benjamins, Coll. Usage-Based Linguistic Informatics, Vol. V, pp. 227-244.

Bacelar do Nascimento, M. F., Gonçalves, J.B., L. Pereira, L., Estrela, A. and Oliveira, S. (2008a), 'Aspectos de unidade e diversidade do português: as variedades africanas face à variedade europeia', Revista Veredas, São Paulo, pp. 35-60.

Bacelar do Nascimento, M. F., Estrela, A., Mendes, A. and Pereira, L. (2008b), 'On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications', Proceedings of the Second Workshop on Building and Using Comparable Corpora (LREC), Marrakech, Morocco, pp. 39-46.

Barreto, F., Branco, A., Ferreira, E., Mendes, A., Bacelar do Nascimento, M.F., Nunes, F. and Silva, J.R. (2006), 'Open resources and tools for the shallow processing of Portuguese: the TagShare project', Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, pp. 1438-1443.

Bettencourt Gonçalves, J. and Veloso, R. (2000), 'Spoken Portuguese: Geographic and Social Varieties', Proceedings of the Second International Conference on Language Resources and Evaluation (LREC), Athens, Greece, Vol. II, pp. 905-908.

Branco, A. and Silva, J. (2004), 'Evaluating Solutions for the Rapid Development of State-of-the-Art POS Taggers for Portuguese', Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Paris, France, pp. 507-510.

Church, K. W. and Hanks, P. (1990), 'Word association norms, mutual information, and lexicography', Computational Linguistics, 16 (1), pp. 22-29.

Cresti, E. and Moneglia, M. (2005) (eds.), C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages, John Benjamins Publishing Company, Studies in Corpus Linguistics nº 15, Amsterdam/Philadelphia (with DVD).

Evert, S. and B. Krenn (2001), 'Methods for the Qualitative Evaluation of Lexical Association Measures', Proceedings of the 39th Meeting of ACL, Toulouse, France, pp. 188-195.

Firth, R. J. (1955), 'Modes of meaning', Papers in Linguistics 1934-1951, London, Oxford University Press, pp. 190-215.

Généreux, M., Hendrickx, I. and Mendes, A. (2012), 'Introducing the Reference Corpus of Contemporary Portuguese Online', Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, pp. 2237-2244.

Marrafa, P., Gonçalves, J.B., Mendes, A. and Veloso, R. (1999), 'A sintaxe do LE-PAROLE', in P. Marrafa, and M. Mota (eds.), Linguística Computacional. Investigação Fundamental e Aplicações, Lisboa, Associação Portuguesa de Linguística/Edições Colibri, pp. 191-205.

Mendes, A., Antunes, S., Bacelar do Nascimento, M.F., Casteleiro, J.M., Pereira, L. and Sá, T. (2006), 'COMBINA-PT: a large corpus-extracted and hand-checked lexical database of Portuguese multiword expressions', Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), Genoa, Italy, pp. 1900-1905.

Mendes, A., Généreux, M., Hendrickx, I., Pereira, L., Bacelar do Nascimento, MF. and Antunes, S. (2012), 'CQPWeb: uma nova plataforma de pesquisa para o CRPC', Textos Seleccionados do XXVII Encontro Nacional da Associação Portuguesa de Linguística, Lisboa.

Pereira, L. and Mendes, A. (2002), 'An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications', Proceedings of the 10th International Congress of EURALEX, Copenhagen, Denmark, vol. II, pp. 841-849.

Sinclair, J. (1991), Corpus, Concordance and Collocation, Oxford University Press, Oxford.

Schmidt, T. (2012), 'EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language', Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey, pp. 236-240.

[i] http://www.clul.ul.pt/index.php.

[ii] http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc.

[iii] http://alfclul.clul.ul.pt/CQPweb/.

[iv] For more information about all the query options, please consult:
http://alfclul.clul.ul.pt/CQPweb/doc/CRPCmanual.v1_en.pdf;
http://alfclul.clul.ul.pt/CQPweb/doc/shortsyntax.v1_en.pdf.

[v] http://www.clul.ul.pt/en/resources.

[vi] https://www.clul.ul.pt/en/resources/ 84-spoken-corpus-qportugues-fundamental-pfq-r.

[vii] http://childes.psy.cmu.edu/manuals/CHAT.pdf.

[viii] http://catalog.elra.info/product_info.php?products_id=1173.

[ix] On April 25th, 1974, a military coup overthrew, with the support of the population, the dictatorial regime, in what is known as the 'Carnation Revolution' (Revolução dos Cravos).

[xi] http://www.clul.ul.pt/en/research-teams/189-c-oral-rom-integrated-reference-corpora-for-spoken-romance-languages.

http://lablita.dit.unifi.it/coralrom.

[xii] http://www.winpitch.com/.

[xiii] http://sites.univ-provence.fr/veronis/logiciels/Contextes/index-fr.html.

[xiv] http://catalog.elra.info/product_info.php?products_id=757.

[xv] http://benjamins.com/#catalog/books/scl.15/main.

[xvi] This corpus was the result of the project entitled Spoken Portuguese – Geographical and Social varieties (more information at http://www.clul.ul.pt/en/research-teams/195-spoken-portuguese-geographical-and-social-varieties).

[xvii] http://www.clul.ul.pt/en/research-teams/ 195-spoken-portuguese-geographical-and-social-varieties.

[xviii] http://catalog.elra.info/product_info.php?products_id=1172.

[xix] http://www.clul.ul.pt/en/research-teams/186-linguistic-resources-for-the-study-of-the-african-varieties-of-portuguese.
http://www.clul.ul.pt/en/research-teams/185-properties-of-african-portuguese-varieties-compared-with-european-portuguese.

[xx] http://www.clul.ul.pt/ en/research-teams/186-linguistic-resources-for-the-study-of-the-african-varieties-of-portuguese.

[xxi] http://www.clul.ul.pt/en/research-teams/197-le-parole.

[xxii] http://catalog.elra.info/product_info.php?products_id=765.

[xxiii] http://cintil.ul.pt.

[xxiv] http://www.clul.ul.pt/en/research-teams/188-tagshare-tagging-and-shallow-morphosyntactic-processing-tools-and-resources.
http://tagshare.di.fc.ul.pt/.

[xxv] The annotation guideline with all the linguistic information encoded in CINTIL is available at: http://nlxserv.di.fc.ul.pt/tagsharecorpus/guidelines.pdf.

[xxvi] http://catalog.elra.info/ product_info.php?products_id=1102.

[xxvii] http://catalog.elra.info/product_info.php?products_id=713.

[xxviii] http://www.clul.ul.pt/en/research-teams/196-simple-semantic-information-for-multifunctional-plurilingual-lexica.

[xxix] For more information on the LE-PAROLE+SIMPLE projects, visit the following website:
http://www.ub.edu/gilcub/SIMPLE/simple.html.

[xxx] http://catalog.elra.info/product_info.php?products_id=713.

[xxxi] http://www.clul.ul.pt/en/research-teams/194-multifunctional-computational-lexicon-of-contemporary-portuguese.

[xxxii] http://www.clul.ul.pt/en/research-teams/187-combina-pt-word-combinations-in-portuguese-language.

[xxxiii] Mutual Information is a lexical association measure, which calculates the frequency of each group in the corpus and contrasts this frequency with the corpus frequency of each word of the group (Church and Hanks, 1990).

[xxxiv] The lexicon and the user manual are available at the project webpage (http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php) and on the Meta-Share repository (http://www.meta-net.eu/meta-share).