The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon

Maria Fernanda Bacelar do Nascimento[1], José Bettencourt Gonçalves[1], Luísa Pereira[1], Antónia Estrela[1], Afonso Pereira[1], Rui Santos[2] and Sancho M. Oliveira[2]

[1]Centro de Linguística da Universidade de Lisboa (CLUL)

[2]Centro de Física Teórica e Computacional, Universidade de Lisboa

Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal

[1]{fbacelar.nascimento, jose.bettencourt, luisa.alice, antonia.estrela}@clul.ul.pt, afonso.pereira@ip.pt

[2]{rsantos, smo}@cii.fc.ul.pt

Abstract

"Linguistic Resources for the Study of the Portuguese African Varieties" is an ongoing project that aims at the constitution, treatment, analysis and availability of a corpus of the African varieties of Portuguese, with 3 million words of written and spoken texts, constituted by five comparable subcorpora, corresponding to the varieties of Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe.

This material will allow intra and intercorpora comparative studies, which will make visible variations that result from discursive and pragmatic differences of each corpus and aspects of linguistic unity or diversity that characterise the spoken Portuguese of this referred five African countries. The five corpora are comparable in size (600,000 words each), in chronology (the last 30 years) and in types and genres (24,000 spoken words and c. 580,000 written words, the last belonging to newspapers, literature and *varia*).

The corpus is automatically annotated and after the extraction of alphabetical lists of lexical forms, these data will be automatically lemmatised. Five separated lists of vocabulary for each variety will be established. A tool for word extraction and preferential calculus according to predefined indexes in order to achieve lexicon comparison of the African Portuguese Varieties is being developed. Concordances extraction will be also performed.

1. Introduction

In the field of language geographical variation, corpora have long been recognised as a valuable source of comparison between language varieties as well as a source for the description of those varieties themselves. Certain corpora have tried to follow, as far as possible, the same sampling procedures as other corpora in order to maximise the degree of comparability. For example, the LOB corpus (The Lancaster - Oslo - Bergen Corpus, compiled at Lancaster (G. Leech), Oslo (S. Johansson) and Bergen (K. Hofland) – 1 million words of British English) contains, roughly, the same genres and sample sizes as the Brown Corpus (compiled by W. N. Francis and H. Kučera at Brown University, Providence - 1 million words of American English) and is sampled from the same year. The Kolhapur Indian corpus is also broadly parallel to Brown and LOB, although the sampling year is different. These three one-million-word corpora have been compared to identify the core vocabulary of international English which is considered to be of fundamental importance for the development of teaching materials and in a linguistic perspective in general (Peyawary, 1999).

Another important project that aims to create comparable corpora of English for use in multiple research and teaching contexts is The International Corpus of English (ICE), compiled by thirteen national groups (including Australia, Canada, East Africa, India, Jamaica, New Zealand, Nigeria, Philippines, UK, USA), coordinated by S. Greenbaum, University College, London.

The corpus CREA (Corpus de Referencia del Español Actual) is also a project that intended to provide the resources for comparative studies of languages used in countries where Spanish is spoken.

All these corpora are available for linguistic description and theory and they reflect the role of natural occurring data by offering a vast program of corpus-based researches as well as opportunities for developing corpus-based language pedagogies.

Given the extreme disparity in the area of Language Resources (LR), regarding, on the one hand, the publication of studies of the European and Brazilian varieties of Portuguese and, on the other hand, the African varieties, the project "Linguistic Resources for the Study of Portuguese African Varieties" aims to fill this gap, providing LR that will allow an objective description of these five African varieties of Portuguese.

Indeed, the lack of empiric data to describe the Portuguese spoken in those countries has been the major concern of researchers, teachers and students of the African countries who have the Portuguese as the official language and, consequently, the establishment of didactic materials.

It is important to say that in the five African countries referred, there is a multilinguistic environment and, in general, the Portuguese is not the mother tongue but only the official language. In Cape Verde, Guinea-Bissau and Sao Tome and Principe, Portuguese based creoles still exist.

The project resulting materials will allow a better description and knowledge of the five African varieties of Portuguese.

2. Constitution of the corpus

The Center of Linguistics of Lisbon University (www.clul.ul.pt) has available a corpus, Corpus de Referência do Português Contemporâneo (CRPC), of 334,711,788 million words including Geographical varieties of Portuguese: Portugal, Brazil, Angola, Cape

Verde, Mozambique, Guinea- Bissau, Sao Tome and Principe, Macao and Goa (represented in different dimensions). The dimension of the written subcorpus is about 332 millions from Books, Newspapers, Magazines, Parliament Sessions, Supreme Court Verdicts, Pamphlets, Correspondence and Miscellaneous including fiction, techno-scientific, didactic and general discourse.

The dimension of the spoken subcorpus is about 2,5 million words and includes informal and formal discourse.

Based in this major corpus, the present work aims at the constitution, treatment, analysis and availability (on-line query) of a corpus of the African varieties of Portuguese, with 3 million words of written and spoken texts, constituted by five comparable subcorpora, corresponding to the varieties of Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe. With the availability of the materials extracted from this corpus, authentic data will be accessible to researchers, teachers, students and authors of different materials (grammars, dictionaries, manuals). These data will be organised making it possible, for the first time, to achieve empirical descriptive studies of each of the Portuguese varieties mentioned above. The same materials will also allow intra and intercorpora comparative studies (of all Portuguese varieties), which will make visible, on the one hand, variations that result from discursive and pragmatic differences of each corpus and, on the other hand, aspects of linguistic unity or diversity that characterise the spoken Portuguese of all five African countries, whose official language is Portuguese. The five corpora are comparable in size (600,000 words each), in chronology (the last 30 years), in types and genres (24,000 spoken words and c. 580,000 written words, the last belonging to newspapers, literature and *varia*).

Thus, our primary goal is to accumulate a body of linguistic knowledge (that would be otherwise difficult to acquire), in order to make possible the use of organised empirical data by academic researchers, students, teachers and material developers, who urgently need to achieve descriptive studies on Portuguese African varieties and to provide comprehensive characterization of lexical and grammatical phenomena, through the study of authentic texts. It will allow the achievement of a global vision of what happens with the Portuguese language when it is used in multilingual contexts, as a foreign language. Furthermore, these corpora will be a valuable source of comparison between these varieties, since corpus data will contain a rich amount of textual information (geographical variety, date, genre, etc.). It will be also possible to detect usages of particular words or phrases as being typical of particular varieties, genres and so on, in order to study the various ways in which lexical and grammatical features occur and recur in actual use.

The CRPC, as a monitor corpus, is the source of the present African Varieties subcorpora. Since the samplings of these varieties are not sufficiently balanced in what concerns their dimension and structure, it was only possible to reach 600.000 words for each variety.

The existing data of African varieties in CRPC include spoken (informal) and written discourse (mainly journalistic, literary and *varia*).

In Table 1, the actual constitution of the African subcorpora in CRPC is presented:

Varieties	Corpus dimension
Angola	c. 15.900.000
Cape Verde	c. 3.200.000
Guinea-Bissau	c. 800.000
Mozambique	c. 2.900.000
Sao Tome and Principe	c. 900.000

Table 1. CRPC African subcorpora

Having in mind the short duration of the project (2 years) and the initial reduced dimension of each subcorpus, the actual corpus structure was designed by evaluating which semantic domains, text genres and periods of time would be easily accessed for the five varieties simultaneously. The structure of similar corpora and the materials already existing where also considered during the corpus design task.

The new materials were collected aiming at the internal balance of each corpus and the compatibility between them. The new parts of the spoken corpus consist of new recordings collected and transcribed by the project's team and the new parts of the written corpus consist of new texts collected either via internet or scanned and reviewed by the project's team.

The balanced corpus that results from this work has the following configuration:

Types and genres		Percentages
Written	**Book**	20% (120.000)
	Newspaper	50% (300.000)
	Varia	26% (156.000)
Spoken	**Spoken**	4% (24.000)
Total		100% (600.000)

Table 2: Percentages of types and genres

3. Annotation and lemmatization

The corpus is automatically annotated and after the extraction of alphabetical lists of lexical forms, these data will be automatically lemmatised. Five separated lists of vocabulary for each variety will be established, according to spoken/written division.

4. Database

In order to make comparisons between the African varieties based in this specific corpus, we prepared and developed a set of informatics tools that will allow to easily treat the data. The set of tools consists mainly on an SQL database together with a number of PERL scripts that simplifies data manipulation.

A search engine will be made available as a set of web pages that will be implemented to interface with the database creating a user friendly environment. It will be possible to introduce in the webpage the words that enable the contextualization of the search. The main goal of the database is the statistical treatment of data, such as word count, frequencies, averages, concordances and others.

The relational database is build with several tables which allow indexation of each word in a text, providing a more effective way to perform concordance extraction. All information related to the texts, like author, editor, etc. is kept, allowing specific analysis of the corpus based on these criteria. We can, for instances, search for statistical information on a word related to a certain author or also for just one or two countries or for a time period.

5. Word and concordances extraction

A tool for words extraction and preferential calculus according to predefined indexes in order to achieve lexicon comparison of the African Portuguese Varieties is being developed. Concordances extraction will be also performed.

The format of the database (machine-readable format easily accessible) creates the necessary conditions for its use, not only by academic researchers but also by other professionals, namely those who work in the educational sector and who have to produce didactic materials for the teaching of Portuguese.

6. Other grammatical and lexical studies

In a second phase of work, we will accomplish the following studies.

6.1 A contrastive study of multiword expression

The objective is to analyse different types of word sequences, from the point of view of the degree of collocational relationships that are established by use, leading to the formal fixedness of the sequence, together with a semantic fixedness; when this process achieves maximum fixedness, the result is a pluriverbal unit, totally lexicalized, i.e., with strong morpho-syntactic and syntactic fixedness (sometimes also phonological) of its elements and with a unitary meaning, memorized as an individual unit.

Despite the difficulty in establishing clear borders between the different types of word combination, this study aims at verifying through quantitative and qualitative analyses which sequences in each variety of African Portuguese present an obvious semantic specialization and a total institutionalization in their use.

6.2 Derivational processes of African Portuguese

The objective is to compare, through regular derivational processes of Portuguese, new lexical forms created in the five varieties and compare these neological forms together with their derivational processes.

6.3 Verbal regencies in African Portuguese

The large differences of verbal regencies of the European Portuguese, the Brazilian Portuguese and the African varieties of Portuguese are well known, but only the exploitation of a corpus with authentic data will clearly show different cases: presence/absence of preposition; differences between used prepositions.

6.4 Loss of casual forms

The objective is to observe the process of loss of casual pronominal forms in a contrastive view, between Portuguese varieties, but also comparing spoken and written corpora. The loss of the accusative and dative pronominal forms in direct and indirect object position and its substitution with a nominative pronominal form is a frequent phenomenon in Brazilian Portuguese, but it is also becoming more usual in European Portuguese. This phenomenon is specially seen in spoken informal discourse. It will be interesting to study this subject also in the African varieties of Portuguese in order to see if there are similarities or not.

6.5 Clitics placement and selection

In what concerns the contrast with the European norm, we can find large differences in the selection of the direct object and indirect object clitics forms. Frequently, instead of the dative form in indirect object context, we can find the accusative form and vice-versa. Also in contrast with the European variety of Portuguese, the collocation of pronouns in pre or post-verbal collocation is different. We aim to study the clitics selection and collocation in the five African varieties of Portuguese.

7. Results

The results of this project will be available at CLUL's webpage for on-line queries, with all the documentation regarding the objectives, methodologies and results. The following materials will be available on-line:

1. Concordances in KWIC format of all the corpus words, organised by subcorpora.
2. Indexation of all the words (lemma and word forms) that occur in the corpus, with frequency data and distribution by subcorpora.
3. Comparative description of the vocabulary of the several subcorpora, as a result of several quantitative and statistic studies.

In this presentation, the following items will be discussed: i) the results accomplished at the time of the meeting, namely the final corpus design with specifications regarding the genres, types and registers of the texts selected for the constitution of the five comparable *corpora*; ii) and partial comparative results of the five lexicons extracted from the corpus under analysis.

8. Acknowledgements

The project "Linguistic Resources for the Study of the Portuguese African Varieties" (PLUS/LIN/50270/2003) has the portuguese funding of Fundação para a Ciência e a Tecnologia (FCT), Gabinete de Relações Internacionais da Ciência e do Ensino Superior (GRICES) and Fundação Calouste Gulbenkian. This project is leaded by Maria Fernanda Bacelar do Nascimento.

9. References

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

Bacelar do Nascimento, M.F. (2002). "Associations lexicales: du corpus aux dictionnaires". In Melka, Francine e Augusto, Maria Celeste (eds.) *De la Lexicologie à la Lexicographie / From Lexicology to Lexicography*, Utrecht Institute of Linguistics (OTS), Utrecht, pp. 38–51.

Bacelar do Nascimento, M.F. (2003). "Quelques considérations sur la constitution et l'exploitation d'un corpus de portugais parlé". In Scarano, A. (a cura di*) Macro-Syntaxe et Pragmatique: L' analyse de l'oral*, Roma: Bulzoni Editore, pp. 295–302.

Bacelar do Nascimento, M.F., Mendes, A. & Pereira, L. (2004). "Providing on-line access to Portuguese language resources: corpora and lexicons", In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (LREC 2004), pp. 1825--1828.

Bacelar do Nascimento, M.F., Mendes, A. & Amaro, R. (2004). "Morphlogical Tagging of a Spoken Portuguese Corpus Using Available Resources", in Branco, A., Mendes, A. & Ribeiro, R. (eds.) *Language Technology for Portuguese: Shallow Processing Tools and Resources.* Lisboa: Colibri, pp. 47--62.

Brito Semedo (1997). *A Colocação dos Clíticos no Português de Maputo*, Maputo, INDE.

Gonçalves, P. & Stroud, C. (org.), (1999). *Panorama do Português Oral de Maputo – Vol. III – Estruturas Gramaticais do Português: Problemas e Exercícios, Maputo*, INDE. Gonçalves, P. & Stroud, C. (org.), (2000). *Panorama do Português Oral de Maputo – Vol. IV –Vocabulário* Básico *do Português (espaço, tempo e quantidade): Contextos e Prática Pedagógica*, Maputo, INDE.

Greenbaum, S. (1990). "The International Corpus of English", ICAME Journal 14: 106--108.

Hofland, K. and S. Johansson (1982). "Word Frequencies in British and American English". Bergen: Norwegian Computing Centre for Humanities.

Leitner, G. (1991). "The Kolhapu corpus of Indian English: intravarietal description and/or intervarietal comparison" in Johansson and Stenström, pp. 215--232.

Miguel, M. H. (2003). *Dinâmica de Pronominalização no Português de Luanda*, Luanda, Editorial Nzila.

Municio, A.M. et al. (2000). "Language Resources Development at the Spanish Royal Academy" LREC proceedings, Athens, Greece, ELRA, pp. 1265--1270.

Peyawary, A.S. (1999). "The Core Vocabulary of International English: A Corpus Approach". Bergen: The Humanities Information Technology Research Programme.

Queffelec, A. (1983). "Inventaire des particularités lexicales du français en Afrique noire, Québec, A.U.P.E.L.F.-A.C.C.T.

Stroud, C. and Gonçalves, P. (org.) (1997). *Panorama do Português Oral de* Maputo *– Vol. II – A Construção de um Banco de "Erros"*, Maputo, INDE.