On the use of comparable corpora of African varieties of Portuguese for linguistic description and teaching/learning applications

Maria Fernanda Bacelar do Nascimento, Antónia Estrela, Amália Mendes, Luísa Pereira

Centro de Linguística da Universidade de Lisboa, University of Lisbon, Portugal

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

www.clul.ul.pt

E-mail: fbacelar.nascimento@gmail.com, antonia.estrela@clul.ul.pt, amalia.mendes@clul.ul.pt, luisa.alice@clul.ul.pt

Abstract

This presentation focuses on the use of five comparable corpora of African varieties of Portuguese (AVP), namely Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe, for multiple contrastive linguistic analyses and for the production of teaching and learning applications. Five contrastive lexicons have been corpus-extracted and further annotated with POS and lemma information and have been crucial to establish for each variety a core and peripheral vocabulary. Studies on AVP-specific morphological processes and on variation in verb complementation will also be discussed. These are first steps towards an integrated description of the five varieties and towards the elaboration of teaching and learning materials to be used by teachers of students from those five African countries with Portuguese as official language.

1. Comparable corpora of African varieties of Portuguese

Compared with the quantity of empirical studies on European Portuguese (EP) and Brazilian Portuguese (BP), developed from corpora and lexicons, the shortage of studies on other varieties of Portuguese is mostly due to the lack of Language Resources (LR).

The Center of Linguistics of the University of Lisbon (CLUL) recently compiled five comparable corpora of the Portuguese Varieties spoken in the five countries which have Portuguese as official language - Angola, Cape Verde, Guinea-Bissau, Mozambique and Sao Tome and Principe. The corpora are available at CLUL's webpage for online query.

The five corpora, which constitute the Africa Corpus, are around 640,000 words each and have the same percentage of spoken and written subparts (c. 25,000 spoken words (4%) and c. 615,000 written words), as shown in Table 1. The written corpus is divided in newspapers (50%), literature (20%) and miscellaneous (26%).

For the task of corpus constitution, some samples of written and spoken materials of already existing corpora compiled at CLUL during the last 30 years were reused, while new recordings were specifically made for this project and new texts were collected.

Countries	Spoken	Written	Total
Angola	27.363	613.495	640.858
Cape Verde	25.413	612.120	637.533
Guinea-Bissau	25.016	615.404	640.420
Mozambique	26.166	615.297	641.463
Sao Tome and Principe	25.287	614.563	639.850
Total	**129.245**	**3.070.879**	**3.200.124**

Table 1. Africa Corpus: constitution and dimension per variety

The five corpora are thus comparable in size, in chronology and in broad types and genres. However, it was not possible to attain comparability at a more granular level. Compiling written materials for each African country proved a difficult task during the project and even more difficult when trying to assure comparable data. Compiling comparable corpora was already a challenge for our group in previous experiences involving European initiatives (like the PAROLE corpora due to the large number of languages involved) and was even more obvious in the case of these African countries. The fact that we only considered texts written by native people living in those five countries made it even more difficult to assure the necessary materials.

Besides the limitation in finding and compiling adequate materials, time was also an important factor, due to the short duration of the project. A follow-up of this work is under way and will assure broader coverage and a more fine-grained comparability of the five corpora.

The newspapers selected are publications with wide national coverage and, regarding fiction, poetry was avoided and only native authors or authors who lived all their lives in the countries were selected. We included in the corpus few texts that are strongly marked, like the case of the African author Mia Couto, whose writings present a high level of lexical creativity and are thus not representative of his AVP.

Since some of the texts collected proved to belong to very different subtypes and genres, it made it difficult to devise specific categories that would accommodate this diversity. This lead us to posit the category "miscellaneous", which corresponds, in fact, to a large collection of heterogeneous texts from different kinds, such as literary or social magazines, computer policies, official documents, religious discourse, political interventions, tourism information, university web pages, academic works, law, national constitution, army information and some short poetry texts. This broad category came to represent a large percentage of the written corpus.

The spoken corpus includes recordings (dialogues and conversations) of spontaneous language on much diversified topics and also recordings from TV and radio

programs. Some previous recordings were used and new ones were made, some by researchers and teachers resident in the five African countries, using their recorder, and some by our own team. The main objective of the recordings was essentially to provide materials for lexical, morphosyntactic and syntactic studies. Since this goal did not require an extreme acoustic quality, the fact that not all recordings were made with high quality equipment was not crucial.

These recordings were orthographically transcribed, following criteria defined according to the project objectives. In what concerns orthographic transcription, in this project, no specific marks for overlaps were used. The orthographic transcription included punctuation signs that usually received the same value they have in writing, but giving special importance to their prosodic marker function, so to transmit, even in a rudimental way, the spoken language rhythm. As a general rule of orthographic transcription, the team transcribed the entire corpus according to the official orthography.

New words following regular patterns of derivation posed no problems for transcription and other cases were transcribed as closely as possible to their pronunciation respecting the Portuguese orthography, and in some cases confirmed with native speakers. Foreign words were transcribed in the original orthography when they were pronounced closely to the original pronunciation. When the foreign words were adapted to the Portuguese pronunciation, they were transcribed according to the entries of the reference dictionaries or according to the orthography adopted in those dictionaries for similar cases. When the speaker mispronounced a word and immediately corrected it, the two spellings were maintained in the transcription of the text. But if the speaker misspelled a word and went on in his speech without any correction, the standard spelling of the word was kept in the transcription. Paralinguistic forms and onomatopoeia not registered in the reference dictionaries were transcribed to represent, as much as possible, the sound produced.

These comparable corpora are the first step towards the development of linguistic studies of the Portuguese Varieties of African countries where Portuguese is the official language and is taught, according to the EP variety, as second or foreign language (Bacelar do Nascimento, 2006; Bacelar do Nascimento *et alii*, 2006 e 2007).

2. Extraction of lexical information

The first studies undertaken based on the five comparable corpora are centered on the contrastive properties of each variety's lexicon: contrastive lexicons of the main POS categories, nuclear *vs*. peripheral vocabulary and divergent derivational processes.

2.1 Corpus annotation

In order to achieve these studies, the five comparable corpora have been automatically annotated with POS and lemma information using Eric Brill's tagger (Brill, 1993), previously trained over a written and spoken Portuguese corpus of 250.000 words, morphosyntactically annotated and manually revised.

The initial tag set for the morphosyntactic annotation of the written corpus covered the main POS categories (Noun, Verb, Adjective, etc.) and secondary ones (tense, conjunction type, proper noun and common noun, variable *vs.* invariable pronouns, auxiliary *vs.* main verbs, etc.), but person, gender and number categories were not included.

2.2 Corpus-extracted contrastive lexicons

Five lexicons had been extracted from the corpora, one per each variety, comprising lexical items from the main categories of Common Name, Adjective and Verb, as well as a category for Foreign Words. For each lexical item, the following information is given: POS, lemma and index of frequency of occurrence in the corpus. A total number of 25.523 lemmas have been described: 14.666 (57%) nouns, 6.268 (25%) adjectives, 4.292 (17%) verbs and 297 (1%) foreign words.

The lexicons of the different varieties have been compared and treated statistically, in the form of contrastive lists, with data of frequency and distribution, and are also available at CLUL's webpage for online query.

2.3 Nuclear *vs*. peripheral vocabulary

One of the most important aspects of the contrastive studies on corpora of varieties of a given language, especially languages such as Portuguese, English, Spanish or French, which are spoken in a great diversity of countries, is to establish the grammatical and vocabulary nucleus to all the varieties. This cohesion will assure the understanding among the speakers of these varieties.

In what concerns English, Quirk *et al.*(1985) agree that: «A common core or nucleus is present in all varieties so that, however esoteric a variety may be, it has running through it a set of grammatical and other characteristics that are present in all the others. It is this fact that justifies the application of the name "English" to all the varieties.» (Quirk et *al.*, 1985, *apud* Nelson, 2006, p. 115).

Using the terminology in Nelson (2006), we have extracted the core vocabulary or nucleus of the five corpora (i.e., the common lexicon to all five varieties), as well as the peripheral vocabulary (i.e., that area of the lexicon where, in the corpus, overlapping between varieties do not occur). The common core data are completely reliable, but even in corpora with bigger dimensions, as the International Corpus of English where each variety is 1M words, it is difficult to consider non-overlapping lexical items as definitively specific of one variety, since many situational and contextual factors may determine the occurrence, or not, of lexical items in one subcorpus. Nevertheless, the results of the peripheral vocabulary must be taken into consideration as being an important contribution to our lexical knowledge of AVP, even though they ought to be validated in corpora of bigger dimensions.

Lexical indexes gave us information on the lemmas that constitute the common nucleus of the five subcorpora and on those that had occurred in four, three, two or only one of the subcorpora. We present in Table 2 the quantitative results, in percentile terms, of these occurrences. As we can see, the percentage of common lemmas to the five corpora is lower than the lemmas that have occurred in only one of the subcorpora.

That common nucleus contains the lemmas with bigger frequency of occurrence in the corpus and it can be considered the Basic Vocabulary of the Africa Corpus.

Core lexicon	Common to 5 varieties	26%
Lexicon From core to periphery	Common to 4 varieties	11%
	Common to 3 varieties	11%
	Common to 2 varieties	15%
Peripheral lexicon	Specific to 1 variety	37%

Table 2. Core and peripheral vocabulary in AVP

This common vocabulary to the five corpora (26% of the lemmas) corresponds to 91.75% of occurrences in the corpus. The lemmas that occurred in just one of the corpora present low frequencies or are hapax legomena and are, in fact, more representative cases of lexical change, or africanization, of the lexicon of the Portuguese language.

2.4 Divergent derivational processes

The neologisms presented in Table 3 were collected in the peripheral zones of the vocabulary and are the result of processes of lexical formation with radicals and affixes available in the European variety. This makes possible morphologic structures that derive from the standards of EP and that, therefore, are predictable and of easy interpretation (Rio-Torto, 2007). We only marked as neologisms lexical items that were not present in the exclusion corpus that we first established (i.e., all the lexical items included in two dictionaries of reference: *Vocabulário da Língua Portuguesa* from Rebelo Gonçalves and *Grande Dicionário da Língua Portuguesa*, Porto Editora) or that were labelled as *africanism*. Of course, this does not mean that some of these neologisms cannot occur in spoken or written productions of EP. We present in Table 3 an example of the lexical productivity encountered in the Africa Corpus, with cases of nouns, verbs and adjectives formed with the prefix *des-* 'un-'.

	Angola	**Cape Verde**	**Guinea-Bissau**	**Mozambique**	**S. Tome and Principe**
Nouns	desatracção desinteriorização	desaculturação descrucificação descravização	desfeitura	descamponês desemergência destriunfo	desarrazoável
Verbs	desconseguir desestrelar	desbaralhar		desconseguir desconter desinventar destrabalhar descosturar desimperializar	
Adjectives	descrispado	desapontador desmamentado		desapetitoso	

Table 3. Neologisms with prefix *des* 'un' in AVP

3. Verb complementation

Verb complementation, at the lexicon-syntax interface, is one of the aspects where AVP are diverging from EP. Based on our preliminary analysis of the five corpora, each AVP shows, in fact, an important internal variation regarding verb complementation (and other properties), either converging or diverging from EP patterns. However, data from the five comparable corpora point to several general tendencies.

3.1 Diverging linguistic properties

First, cases where direct objects in AVP (corpus examples (1a) and (2a)) occur as indirect or prepositional objects in EP ((1b) and (2b)):

(1)a. "Pediram o Ministério da Educação Nacional para assumir a sua responsabilidade" G(W)[1]
'[They] asked the Ministry of National Education-dirOBJ to assume its responsibility'

b. Pediram ao Ministério da Educação Nacional que assumisse a sua responsabilidade (EP)
'[They] asked to the Ministry of National Education-indirOBJ that it assume its responsibility-dirOBJ

(2)a. "tinha que ir a escola" M(S)
'[I] had to go the school-dirOBJ

b. tinha que ir à escola
'[I] had to go to the school-prepOBJ

Second, the opposite situation where indirect or prepositional objects in AVP ((3)-(4)) correspond in EP to direct objects:

(3)a. "o Adolfo pegou então a doença que lhe foi matar" A(W)
'Adolfo got then the disease that him-indOBJ killed (killed him)

b. o Adolfo apanhou então a doença que o foi

[1] The codes following the corpus examples indicate their origin: A – Angola; CV – Cape Verde; G – Guinea-Bissau; M – Mozambique; ST – Sao Tome and Principe; S – Spoken; W – Written. The same codes are used for countries in Tables 4-7.

matar (EP)
'Adolfo got then the disease that him-dirOBJ killed

(4)a. "para combater com a delinquência" G(S)
'to fight with the delinquency-prepOBJ'

b. para combater a deliquência (EP)
'to fight the delinquency-dirOBJ'

And third, the fact that complements are frequently introduced in AVP by a different preposition than the one occurring in EP, like in (5). It seems that in AVP the range of prepositions tends to be more limited and some prepositions are extensively used in contexts which would show in EP a large variation, like the case of preposition *em* 'in' (see example (5)) which covers different semantic values.

(5)a. "Menino esperto, você precisa ir na escola" A(W)
'Smart boy, you need to go in-the school

b. Menino esperto, você precisa de ir à escola (EP)
'Smart boy, you need to go to-the school

Moreover, pronominal verbs in EP, either intrinsic pronominal verbs or verbs intrinsically pronominal in a specific meaning, do occur very frequently in AVP as non-pronominal:

(6)a. "O Partido da Renovação Social (PRS) congratulou ontem com a nomeação de Aristides Gomes" G(W)
'The Party of the Social Renovation (PRS) congratulated [himself] yesterday with the nomination of Aristides Gomes'

b. O Partido da Renovação Social (PRS) congratulou-se ontem com a nomeação de Aristides Gomes (EP)
'The Party of the Social Renovation (PRS) congratulated-clitic=himself) yesterday with the nomination of Aristides Gomes'

This affects furthermore inchoative alternations, typically pronominal with certain lexical verb classes in EP, and frequently lexicalized as non-pronominal in AVP, like in (7), a behaviour that requires an in-depth contrastive analysis of lexical verb classes in AVP so as to capture relevant insights on the meaning-syntax relationship.

(7)a. "O seu partido não preocupa com o abandono ou não de Tagme Na Waye" G(W)
'His party does not worry with the abandonment or not of Tagme Na Waye'

b. O seu partido não se preocupa com o abandono ou não de Tagme Na Waye (EP)
'His party does not se-clitic=himself worry with the abandonment or not of Tagme Na Waye'

The differences in verb complementation presented above imply significant changes in the syntax of AVP. For example, the lexicalization of indirect objects as direct objects leads to a structure with double objects, a possibility excluded in EP (see example (8)).

(8)a. "perguntas uma pessoa 'o que é que tu queres fazer?" G(S)
'[you] ask a person-dirOBJ 'what do you want to do'-dirOBJ'

b. "perguntas a uma pessoa 'o que é que tu queres fazer?" (EP)
'[you] ask to a person-IndirOBJ 'what do you want to do'-dirOBJ'

The general tendency to transform indirect and prepositional objects into direct objects leads to the possibility of forming structures like passive and certain inchoative alternations with verbs which do not allow for those constructions in EP (see passive construction in (9)):

(9)a. "É-lhe informado que a chegada do barco seria no dia seguinte" CV(W)
It was-to him-indOBJ informed that the arrival of the boat would be on the next day-SUBJ
'He was informed that the arrival of the boat would be on the next day'

b. Ele é informado de que a chegada do barco seria no dia seguinte (EP)
He-SUBJ was informed of that the arrival of the boat would be on the next day

The passive construction reveals other important aspects, namely, the fact that passives are encountered in AVP with verb classes which do not allow passivization in EP. It is the case of the verb *nascer* 'to be born' (10), an unaccusative verb, i.e. a verb with a subject argument that presents certain syntactic and semantic properties that differ from typical intransitive verbs.

(10)a. "em sessenta e seis fui nascido" A(S)
'in sixty six [I] was born'

b. em sessenta e seis nasci (EP)
in sixty six [I] born

3.2 Analysis of specific lemmas

Although the properties sketched above are pervasive in the five AVP corpora, we wanted to observe possible differences in verb complementation in each variety. In order to do so, we first started by analysing, in the five comparable corpora, concordances of verb lemmas with different syntactic structures, belonging to different lexical classes and in some cases, having a pronominal construction or participating in a pronominal inchoative alternation: *matar* 'to kill', *responsabilizar* 'to hold responsible / to assume responsability', *informar* 'to inform', *combater* 'to fight', *perguntar* 'to ask a question', *pedir* 'to ask for', *habituar* 'to make/get used to', *chegar* 'to arrive', *voltar* 'to return', *precisar* 'to need', *congratular* 'to congratulate', and *preocupar* 'to worry'.

The general results are presented in Table 4, with information, for each variety, on the type of linguistic phenomena, the number of contexts showing divergence

from EP (DF), the total frequency of the lemmas considered for this study (TF) and the percentage of diverging contexts from EP. The general phenomena considered are the passage of direct objects to indirect objects (DO > IO) or to prepositional objects (DO > PP), the opposite, namely the passage of indirect or prepositional objects to direct objects (IO/PP > DO), the use of a different preposition introducing a prepositional object, the absence of preposition *de* 'of' introducing noun phrases or clauses and the use of pronominal constructions (either reflexive, inherent or anticausative) as non pronominal ones.

The first important information to draw from these results is the fact that contexts diverging from EP in the corpora are not largely extensive and that the lemmas behave in most cases according to the European norm.

Variety ►	**A**			**CV**			**G**			**M**			**ST**		
Linguistic phenomena ▼	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**
DO > IO	4	250	**1,6**	-	185	**0,00**	12	225	**5,33**	8	376	**2,13**	1	250	**0,40**
DO > PP	-	32	**0,00**	-	35	**0,00**	1	102	**0,98**	0	51	**0,00**	-	33	**0,00**
IO / PP > DO	1	349	**0,29**	1	408	**0,25**	11	426	**2,58**	3	354	**0,85**	3	216	**1,39**
different preposition	26	1163	**2,24**	2	946	**0,21**	5	1700	**0,29**	6	872	**0,69**	8	644	**1,24**
no preposition DE	19	197	**9,64**	54	246	**21,95**	21	270	**7,78**	23	181	**12,71**	21	174	**12,07**
pron. > non pron.	1	137	**0,73**	-	116	**0,00**	13	170	**7,65**	1	102	**0,98**	3	92	**3,26**
Total	51			57			63			41			36		

Table 4. Diverging syntactic behaviour regarding EP of selected lemmas in the 5 AVP

The second aspect is that these data confirm our first impression when confronted with the five AVP corpora, namely the fact that the Portuguese variety of Guinea-Bissau is the one presenting more diverging patterns regarding EP, although, of course, results are still preliminary. In fact, in Table 4, the Portuguese of Guinea-Bissau presents the most diverging numbers in comparison with EP in what concerns direct objects (DO) becoming indirect objects (IO) or prepositional objects (PP), indirect objects or prepositional objects realized as direct objects, and, also, the use of pronominal verbs as non-pronominal (pron. > non pron.). The opposite general tendency occurs with the variety of Cape Verde, with almost no verbal contexts differing from EP. But surprisingly, the Cape Verde variety shows an extremely high number of occurrences where preposition *de* is omitted. Although the absence of preposition *de* introducing noun phrases or clauses is also a general tendency in EP, the large majority of the cases encountered in AVP differs from the ones found in EP. Table 4 points to several linguistic phenomena where all five varieties diverge from the pattern of EP (the transitivization of verbs, the change of preposition introducing verb complements and the absence of preposition *de* introducing object noun phrases and clauses) while two other patterns occur in 4 varieties (direct objects as indirect ones and the realization of pronominal constructions as non pronominal). These data point to the fact that at least four AVP share a general tendency towards changes in verb complementation, leaving the Cape Verde variety as a special case where a specific tendency is uncovered.

	A			**CV**			**G**			**M**			**ST**		
	DF	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**	**DF**	**TF**	**%**
Different preposition	**26**	**1163**	**2,24**	**2**	**946**	**0,21**	**5**	**1700**	**0,29**	**6**	**872**	**0,69**	**8**	**644**	**1,24**
Perguntar *OI > PP*	8	161	**4,97**	-	177	**-**	-	147	**-**	-	149	**-**	-	37	**-**
responsabilizar *por* 'by' > *de* 'of'	1	15	**6,67**	-	13	**-**	4	34	**11,76**	-	19	**-**	-	18	**-**
Responsabilizar *por* 'by' > *em* 'in'	-	15	**-**	-	13	**-**	1	34	**0,68**	-	19	**-**	1	18	**2,70**
chegar *a* 'to' > *em* 'in'	12	441	**2,72**	2	449	**0,45**	1	458	**0,22**	5	376	**1,33**	7	317	**2,21**
voltar *a / para* 'to' > *em* 'in'	5	523	**0,96**	-	279	**-**	-	1028	**-**	-	319	**-**	-	256	**-**
habituar *a* 'to' > *com* 'with'	-	23	**-**	-	28	**-**	-	33	**-**	-	9	**-**	1	16	**6,25**

Table 5. Detailed syntactic behaviour of a specific linguistic property in AVP

Although these conclusions are certainly accurate in what regards the corpus data which was analysed, when observing the results in more detail, one is confronted with a larger amount of variation for the set of lemmas under study among the five varieties. In Table 5, one of the patterns mentioned in Table 4, namely the tendency for the change of

preposition introducing object noun phrases and clauses, is furthermore detailed into the lemmas presenting contexts diverging from EP. Although change in preposition is a property very generally attributed to AVP and although this is a property pervasive of all five varieties, we see that frequencies are in fact low, with some more cases in Angola, but with the rest of the AVP presenting some lemmas with just one diverging context. So, the rising percentages correspond in many cases to non significant context frequencies. When observing each lemma independently in each variety, we see that they do behave in very different ways. While *chegar em* 'to arrive in' is a systematic verb complementation change in all AVP, *perguntar em* 'to ask in' is a somehow frequent pattern only in Angola. The general patterns of change in verb complementation only arise in Table 5 and in the full results when grouping contexts from different lemmas.

These data strongly confirm the need to treat each AVP as an independent system showing specific tendencies, even if the five properties do share more general principles of change. Considering our objective of preparing materials for the teaching and learning of Portuguese in those five African countries, we soon understood the need to establish different manuals for each variety, focusing on the diverging phenomena regarding EP, uncovered by corpus analysis. For example, in the case of Guinea-Bissau, all the linguistic aspects analysed in this corpus would have to be covered, but taking into attention specific lemmas which proved to present more diverging patterns from the EP norm, like the verbs *informar*, *precisar* and *preocupar*. While the Portuguese variety of Angola will need more focus on the change of preposition with verb *perguntar* and *responsabilizar* and the Cape Verde variety will essentially need some attention to the absence of preposition *de*, especially with verb *precisar* when followed by an infinitive verb.

The choice of four lemmas (verbs *preocupar*, *responsabilizar*, *congratular* and *habituar*) pointed to the fact that pronominal constructions tend to be used as non pronominal, and the global frequencies were presented in Table 4. But a closer look to other verbs (non pronominal in EP) show contexts where those are used pronominally, with a clitic pronoun which does not correspond to a verb complement, but rather to a particle with different functionalities. This raises the question of whether we are facing a tendency towards a non pronominal use of pronominal verbs or a mixed tendency, towards both insertion and loss of clitic pronoun. We then analysed all the contexts where the two patterns were found in the five spoken subcorpora, for all lemmas. The results are presented in Table 6, with information, for each variety, on the total frequency of contexts showing insertion or absence of a clitic pronoun and thus diverging from EP.

Pronoun	**A**	**CV**	**G**	**M**	**ST**
Insertion	4	5	11	9	5
Absence	23	29	63	42	75

Table 6. Differences in pronominal constructions in AVP compared to EP

Indeed, numbers are clearly more significant in the case of non pronominal uses of verbs which would be pronominal in EP, and this tendency is shared by all AVP. The insertion of a clitic pronoun in contexts of non pronominal constructions could be seen as a consequence of the tendency to omit the pronoun, since this would inevitably generate some confusion on the use of pronominal constructions and also a tendency towards overcorrection.

It is important to take into account that we are facing varieties of Portuguese that have not yet reached a stable point of evolution, so that only a detailed contrastive analysis of more data concerning verb complementation could point us towards the path of understanding the current ongoing changes.

4. Degree of variation regarding European Portuguese

As already mentioned, most of the verb lemmas that we analysed in the corpus had low frequencies and especially low frequencies of contexts diverging from EP. Another question was the strong variation found among the five AVP regarding those lemmas, which seemed to make it difficult to assess the degree by which each AVP differ from the European norm. However, when observing texts from the different varieties, the higher or lower degree of variation was more evident. This lead us to search for the totality of diverging contexts regarding some linguistic phenomena instead of focusing on some lemmas. Of course, this objective was not doable over the whole corpus and we decided to limit this study to the spoken subcorpus, since it does present more AVP-specific patterns than the written one.

Four important linguistic phenomena showing divergence from the EP standard were selected for a comparison between the five spoken corpora: the position of the clitic pronoun regarding the verb, the concordance inside the nominal phrase and between subject and predicate, verbal conjugation and insertion or absence of clitic pronouns (already discussed in Table 6). The results are presented in Table 7.

	A	**CV**	**G**	**M**	**ST**
Position of clitic pronouns	**40**	**23**	**46**	**68**	**20**
Concordance: total	**136**	**57**	**241**	**116**	**64**
Nominal concordance					
Determiner-noun: gender	6	5	47	12	4
Determiner-noun: number	38	5	39	27	11
Noun-adjective: gender	2	5	22	2	2
Noun-adjective: number	13	6	21	11	1
Subject-predicative noun: gender	13	-	16	6	5
Subject-predicative noun: number	13	-	2	4	3
Verbal concordance					
Person	25	10	39	8	5
Number	26	26	55	46	33
Conjugation	**59**	**29**	**96**	**78**	**43**
Mood	26	19	69	52	20
Time	33	10	27	26	23
Pronominal constructions	**27**	**34**	**74**	**51**	**80**
Insertion of clitic pronoun	4	5	11	9	5
Absence of clitic pronoun	23	29	63	42	75
Total	**262**	**143**	**457**	**313**	**207**

Table 7. Patterns of variation in the five AVP

Table 7 includes the number of diverging occurrences in each variety and for each property, together with the final number of diverging contexts for each country. These contexts were found over a total number of around 25-27 thousand words, the number of words of each spoken subcorpus.

When looking globally at different linguistic aspects in the whole subcorpora and comparing the data in Table 7, the variety of Guinea-Bissau emerges as the most diverging one regarding EP. It comes as a confirmation of the conclusion already attained with the study of verb complementation (which seemed however limited by low frequencies). Although the Mozambican variety shows higher results in what concerns the position of clitic pronouns and although the variety of Cape Verde shows slightly higher results in what concerns pronominal constructions, the contexts regarding concordance and verb conjugation and the global results isolate Guinea-Bissau. The other varieties follow gradually, Mozambique, Angola, Sao Tome and Principe, in that order, and, finally, Cape Verde, which is the less diverging variety of the five, as the results of verb complementation already showed.

4. Conclusions

If, on the one hand, the adequate description of the linguistic properties of AVP requires the use of balanced corpora, it is also true that, on the other hand, the evaluation of the degree by which they diverge from the EP norm as well as the contrastive study the five AVP will only be possible through the access to comparable corpora of those varieties.

The recently compiled comparable corpora of the AVP are the first step towards a better understanding of the similarities and differences encountered among them and EP and between each variety. The first linguistic results based on these corpora have been a contrastive lexicon which establishes a common core vocabulary, as well as peripheral lexical sets for each variety. Two strongly diverging phenomena are under contrastive study, morphological and lexical analysis of derivational processes in AVP, as well as verb complementation.

However important the results may be, in order to reach confident observations regarding the evolution of AVP and their relationship with EP, it is necessary to ensure the enlargement of the existing five comparable corpora, with special attention to the spoken subpart, since most AVP-specific properties only show in the spoken register. In fact, since most of the properties where AVP differs from the EP norm are still emergent and show strong variation inside each variety, it is essential to rely on comparable corpora which are balanced. Only then will we be able to establish more stable tendencies of linguistic change across the varieties and inside each variety.

This will be important to give teachers more knowledge about the linguistic properties that are characteristic of the African varieties of Portuguese as well as teaching and learning materials that could point to the process of identification and understanding of the unity and diversity factors that are at stake between these varieties and between those and European Portuguese.

5. Aknowledgements

The Corpus Africa was compiled and explored in the scope of the projects "Linguistic Resources for the Study of the African Varieties of the Portuguese", of the Center of Linguistics of the University of Lisbon (CLUL) and of the Center of Theoretical and Computational Physics of the University of Lisbon, and "Properties of the African Varieties compared with the European Portuguese". The first project had been financed by the Foundation for Science and the Technology under the cover of the Lusitânia Program and for the Foundation Calouste Gulbenkian and the second had been financed only by this Foundation. The first project was supported by Center de Information and Documentation Amílcar Cabral (CIDAC), the Open University of Lisbon, the Camões Institute, the Embassy of the Democratic Republic of S. Tome and Principe and the Group of Natural Language of the Department of Computer Science of the College of Sciences of the University of Lisbon. The consultant of the first project was Perpétua Gonçalves and, of the second, Perpétua Gonçalves and Amália Mendes (For more information on the projects, see the webpage http://www.clul.ul.pt/sectores/linguistica_de_corpus/linguistica_de_corpus.php.

6. References

Bacelar do Nascimento, M. F. , Bettencourt Gonçalves, J. (2006). The Role of Spoken Corpora in Teaching/Learning Portuguese as a Foreign Language - The Case of Adjectives of Intensification. In Kawaguchi, Y., Zaima, S., Takagaki T. (eds.) *Spoken Language Corpus and Linguistic Informatics*. Amsterdam: Jonh Benjamins, Coll. Usage-Based Linguistic Informatics, vol.V, pp. 219-226.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela , A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2006). The African Varieties of Portuguese: Compiling Comparable Corpora and Analyzing Data-derived Lexicon". In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC2006)*, 24-26 May, Genoa, Italy, pp. 1791-1794.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela, A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2007). As variedades africanas do português: um corpus comparável". *X Simposio Internacional de Comunicación Social*, Ministerio de Ciencia, Tecnología y Medio Ambiente, v. I, Janeiro, Santiago de Cuba.

Bacelar do Nascimento, M. F., Pereira, L. A. S., Estrela, A., Bettencourt Gonçalves, J., Oliveira, S. M., Santos, R. (2007). Especificidades das Variedades Africanas do Português na Formação dos Professores de Português". *Saber Ouvir, Saber Falar*, *7º Encontro Nacional da Associação de Professores de Português* (CD-ROM).

Gonçalves, P., Stroud, C. (Orgs.) (1998). *Panorama do Português Oral de Maputo*, Vol. 3 – *Estruturas Gramaticais do Português: Problemas e Exercícios*, Cadernos de Pesquisa nº 27. Maputo, INDE.

Gonçalves, P., Stroud, C. (Orgs.) (2000). *Panorama do Português Oral de Maputo*, Vol. 4 – *Vocabulário Básico do Português (espaço, tempo e quantidade) Contextos e Prática Pedagógica*, Cadernos de Pesquisa nº 36. Maputo, INDE.

Gonçalves, P., Stroud, C. (Orgs.) (2002). *Panorama do Português Oral de Maputo*, Vol. 5 – *Vocabulário Básico do Português, Dicionário de Regências*, Cadernos de Pesquisa nº 41. Maputo, INDE.

Gonçalves, P. (1997). Tipologia de Erros. In Stroud, C., Gonçalves, P. (Orgs.) (1997b).

Gonçalves, R. (1966). *Vocabulário da Língua Portuguesa*. Coimbra: Coimbra Editora.

Grande Dicionário da Língua Portuguesa (2004). Porto, Porto Editora.

Greenbaum, S. (1996). *Comparing English Worldwide: The International Corpus of English*. Oxford, Clarendon Press.

Nelson, G. (2006). The core and the periphery of world Englishes: a corpus-based exploration. *World Englishes*, 25:1, pp. 115-129.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London, Longman.

Rio-Torto, G. (2007). Caminhos de Renovação Lexical: Fronteiras do Possível. In Isquerdo, A. N., Ieda, M. A. (Orgs.), *As Ciências do Léxico, Lexicologia, Lexicografia, Terminologia*, Vol. 3. Campo Grande, MS, Ed. UFMS; São Paulo, Humanitas.

Stroud, C., Gonçalves, P. (Orgs.) (1997a). *Panorama do Português Oral de Maputo*, Vol. 1 – *Objectivos e Métodos*, Cadernos de Pesquisa nº 22. Maputo, INDE.

Stroud, C., Gonçalves, P. (Orgs.) (1997b). *Panorama do Português Oral de Maputo*, Vol. 2 – *A Construção de um Banco de " Erros"*, Cadernos de Pesquisa nº 24. Maputo, INDE.