CORPORA COMPARÁVEIS E VARIAÇÃO LEXICAL NAS VARIEDADES AFRICANAS DO PORTUGUÊS

Maria Fernanda BACELAR DO NASCIMENTO[1]

- RESUMO: Neste artigo são apresentados resultados de um projecto de constituição de *corpora* orais e escritos das cinco variedades africanas do português (Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe) e de extracção dos respectivos léxicos. Os cinco *corpora*, que no total perfazem 3.200.124 palavras, são comparáveis em dimensão, constituição interna e cronologia. Com a realização deste trabalho, pretendeu-se dar uma contribuição para colmatar uma grave lacuna no que respeita aos Recursos Linguísticos (*corpora* e léxicos) existentes para o português. Existiam já, em elevado número e grandes dimensões, *corpora* e léxicos das variedades europeia e brasileira do português, mas eram quase inexistentes (se excluirmos Moçambique) Recursos Linguísticos semelhantes para as variedades africanas que permitissem análises objectivas de cada uma delas e que, pela sua comparabilidade permitissem estudos contrastivos entre essas variedades ou entre elas e o português europeu ou do Brasil. O projecto, intitulado Recursos Linguísticos para o Estudo das Variedades Africanas do Português, foi executado pelo grupo Linguística de *Corpus* do Centro de Linguística da Universidade de Lisboa – CLUL e por uma equipa do Centro de Física Teórica e Computacional da mesma universidade, tendo sido acompanhado pela consultora do projecto, Perpétua Gonçalves, da Universidade moçambicana Eduardo Mondlane.

- PALAVRAS-CHAVE: *Corpus* África; léxico; variação; português.

Introdução

Nas últimas décadas tem sido crescente o interesse pelo estudo dos usos da linguagem e da variação e mudança linguística com base em *corpora*. No âmbito desses estudos, a constituição de *corpora* comparáveis tem propiciado a realização de análises contrastivas entre produções linguísticas de falantes de variedades de uma mesma língua que a utilizam quer como língua materna quer como língua segunda. Os *corpora* constituídos com esta finaliddae contêm, em geral, amostragens de diversos registos e níveis de língua e de vários géneros e

[1] Universidade de Lisboa – Centro de Linguística – 1649-003 – Lisboa – Portugal. Endereço eletrônico: fbacelar.nascimento@clul.ul.pt

tipos de discurso. Considerando sempre factores contextuais no estudo da variação linguística, têm sido realizadas sobre estes *corpora* análises a nível lexical, morfológico, morfo-sintáctico, sintáctico, semântico, pragmático ou estilístico (REPPEN; FITZMAURICE; BIBER, 2002).

Os objectivos destes estudos são, essencialmente, descrever objectivamente as variedades observadas através da análise empírica de dados naturais, compará-las quantitativamente (a partir de resultados de Frequência alinhados) ou qualitativamente (a partir de unidades lexicais ou multilexicais ou de estruturas sintácticas contextualizadas e alinhadas) e, finalmente, obter informações comparativas. Estas informações permitem identificar variantes históricas, estabelecer correlações entre fenómenos, verificar mudanças (ou tendências de mudança) sintácticas e semânticas e, ainda, calcular diversos graus de variação consoante os registos de língua (oral ou escrito), os níveis de língua (formal ou informal) e os géneros e tipos de discurso (literário, jornalístico, académico, técnico ou científico, etc.).

Alguns autores de *corpora* têm procurado seguir, tanto quanto possível, os mesmos procedimentos de constituição das amostragens usados em *corpora* já existentes, com o fim de maximizar o grau de comparabilidade dos dados. Por exemplo, o *The Lancaster – Oslo – Bergen Corpus* (LOB *Corpus*), compilado por equipas de investigadores em Lancaster, sob a coordenação de G. Leech; em Oslo, sob a coordenação de S. Johansson; e em Bergen, sob a coordenação de K. Hoffland) com 1 milhão de palavras de inglês britânico, contém, aproximadamente, os mesmos géneros e dimensões de amostragens do Brown *Corpus* (compilado por W. N. Francis e H. Kucera na Brown University, Providence), também este com 1 milhão de palavras mas do inglês da América e a mesma cronologia. O *Kolhapur Indian Corpus* é também largamente comparável aos *corpora* Brown e LOB, embora a recolha recaia sobre um período de tempo diferente. Estes três *corpora*, de 1 milhão de palavras cada, têm sido comparados para, entre outros estudos, identificar um núcleo vocabular do chamado International English, considerado de importância fundamental numa perspectiva linguística geral e, particularmente, para o desenvolvimento de materiais para o ensino da língua inglesa (PEYAWARY, 1999). Outro importante projecto que visa a criação de *corpora* comparáveis do inglês para uso em múltiplas investigações e contextos de ensino é *The International Corpus of English* (ICE), compilado por treze grupos nacionais (incluindo Austrália, Canadá, África Oriental, Índia, Jamaica, Nova Zelândia, Nigéria, Filipinas, Reino Unido e Estados Unidos da América) sob a coordenação de S. Greenbaum do University College de Londres.

O *Corpus de Referencia del Español Actual* (CREA) é outro projecto que pretende providenciar recursos para estudos comparativos de variedades do espanhol.

Para além dos estudos lexicais, há inúmeros exemplos de estudos gramaticais resultantes de investigação realizada sobre *corpora* de variedades de uma mesma língua dos quais é de salientar *The Longman Grammar of Spoken and Written English* (BIBER, D. et al., 1999) que adopta uma abordagem baseada num *corpus* do inglês britânico e do inglês da América.

Em Portugal, o *Corpus* de Referência do Português Contemporâneo (CRPC), do Centro de Linguística da Universidade de Lisboa (CLUL), coordenado por Maria Fernanda Bacelar do Nascimento, contém, actualmente, mais de 300 milhões de palavras e inclui todas as variedades do português. Trata-se de um *corpus* monitor, isto é, de um *corpus* em que vão sendo incluídos todos os documentos a que a equipa vai tendo acesso, sem preocupação de equilíbrio interno do *corpus*. Recorre-se a este para a constituição dos *corpora* de base de diversos trabalhos, *corpora* que são desenhados e construídos tendo em conta os objectivos visados pelos trabalhos em causa.

Objectivos do trabalho

Tendo em consideração a extrema desigualdade que se verifica, no que respeita a Recursos Linguísticos e à publicação de estudos, entre, por um lado, as variedades europeia e brasileira do português e, por outro lado, as variedades africanas (excluindo os *corpora* orais de Moçambique e os valiosos estudos sobre eles publicados), o grupo Linguística de *Corpus* do CLUL decidiu levar a cabo um projecto de trabalho que teve como principal objectivo preencher essa lacuna, fornecendo Recursos Linguísticos comparáveis que possibilitam descrições objectivas das cinco variedades africanas e estudos contrastivos entre essas variedades ou entre elas e o português europeu e do Brasil.

O projecto realizado intitula-se "Recursos Linguísticos para o Estudo das Variedades Africanas do Português", teve duração de 28 meses e ficou concluído em Dezembro de 2006.[2]

Deste projecto resultou o *Corpus* África, com a dimensão de 3.200.124

[2] O projecto foi executado pelo Centro de Linguística da Universidade de Lisboa em parceria com o Centro de Física Teórica e Computacional da mesma Universidade. Coordenado por Maria Fernanda Bacelar do Nascimento, foi realizado por duas equipas constituídas pelos seguintes investigadores: do Centro de Linguística, Luísa Alice Santos Pereira, Antónia Estrela, José Bettencourt Gonçalves e Afonso Pereira; do Centro de Física Teórica e Computacional , Rui Santos e Sancho Oliveira. Foi consultora do projecto Perpétua Gonçalves, da Universidade Eduardo Mondlane de Maputo. O projecto foi financiado pela Fundação para a Ciência e a Tecnologia através do Programa Lusitânia, pelo Serviço de Educação e Bolsas da Fundação Calouste Gulbenkian e pelo Gabinete de Relações Internacionais da Ciência e do Ensino Superior. Teve, ainda, o apoio do Centro de Informação e Documentação Amílcar Cabral (CIDAC), do Instituto Camões, da Embaixada de São Tomé e Príncipe, da Universidade Aberta e do Grupo de Fala e Linguagem Natural (NLX) do Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa.

palavras,[3] que inclui os cinco *sub-corpora* de Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe, um vocabulário constituído por palavras plenas extraídas daqueles *sub-corpora* e indexadas contrastivamente e a obtenção de concordâncias, em formato KWIC, de todas as ocorrências do *corpus*, organizadas por *subcorpora* e por registo oral e escrito.

Constituição do *Corpus* África

O CRPC continha, à data do início dos trabalhos, um conjunto substancial de textos de língua portuguesa em África com mais de 20 milhões de palavras. Contudo, o requisito, essencial ao projecto, de comparabilidade dos dados exigia que os cinco *subcorpora* tivessem dimensão e estrutura interna semelhantes, o que não permitiu que o *Corpus* África fosse constituído recorrendo apenas aos materiais do CRPC. Nalguns casos foi possível extrair deste *corpus* monitor uma significativa quantidade de textos; no entanto, ele não continha documentos equivalentes, em todas as variedades africanas em observação, para os registos e os géneros previstos aquando do desenho equilibrado do *Corpus* África. Foi assim para muitos casos, de que são exemplo os textos que constituem os *corpora* orais, muito escassos para quase todas as variedades ou, ainda a título de exemplo, a quase inexistência no CRPC, de textos literários de São Tomé e Príncipe.

Assim, depois de seleccionados os materiais existentes que eram compatíveis com os objectivos do projecto, procedeu-se a um intenso trabalho de recolha de novos materiais quer para o *corpus* oral (gravação, transcrição e tratamento de discurso oral), quer para o *corpus* escrito (selecção de livros, jornais, revistas e outros documentos em bibliotecas e via internet e tratamento dos textos), a que se seguiu a informatização e codificação dos documentos e a sua inclusão numa base de dados.

A estrutura interna do *Corpus* África é a seguinte:

- *corpus* escrito – 96%, assim distribuído:

livro literário	20%
jornal	50%
varia	26%

- *corpus* oral – 4%

[3] Na proposta de projecto inicialmente apresentada, previa-se que o *corpus* fosse constituído por 5 milhões de palavras, 1 milhão por cada *sub-corpus*. A escassez do financiamento atribuído através do Programa Lusitânia, embora em parte colmatada por um subsídio posteriormente atribuído ao projecto pela Fundação Calouste Gulbenkian, obrigou à redução do *corpus* projectado.

O *corpus* oral é essencialmente constituído por discurso informal – conversas espontâneas mas também por algum discurso formal – entrevistas de rádio e discursos políticos. Contém transcrições de 80 gravações, 45 de homens e 35 de mulheres. Do total deste *corpus*, 80% dos informantes tem um nível de escolaridade médio ou superior e 20% um nível de escolaridade primário.

Dadas as limitações financeiras a que o projecto esteve sujeito, este foi o *corpus* possível. Teria sido importante atingir 1 milhão de palavras para cada variedade (como estava previsto no programa inicial) e dar mais peso ao *corpus* oral que ficou mais prejudicado pois, como é sabido, implica mais custos em tempo e em recursos humanos.

As 3.200.124 palavras estão distribuídas no *corpus* da seguinte forma:

PAÍSES	*CORPUS* ORAL	*CORPUS* ESCRITO
ANGOLA	27.363	613.495
CABO VERDE	25.413	612.120
GUINÉ	25.016	615.404
MOÇAMBIQUE	26.166	615.297
SÃO TOMÉ E PRÍNCIPE	25.287	614.563
TOTAL	129.245	3.070.879
TOTAL DOS DOIS *CORPORA*	3.200.124	

Quadro 1 – Dimensão e distribuição do *corpus* África

Anotação do *corpus*

Foram testados dois anotadores, tendo-se concluído pela maior eficácia, no que respeita à percentagem de erro, do anotador Eric Brill (http://www.cs.jhu.edu/~brill) adaptado ao português europeu pelo grupo Linguística de *Corpus* do CLUL. Este anotador especialmente feito para *corpora* escritos e que a equipa mais tarde adaptou também à anotação de *corpora* orais, no âmbito do projecto C-ORAL-ROM, destina-se a anotação morfo-sintáctica e cobre as principais categorias constantes das Partes do Discurso (Nome, Verbo, Adjectivo, Advérbio, Pronome, Preposição, Conjunção, etc.) e categorias secundárias (modo e tempo verbal, nome próprio e comum, tipos de pronomes e conjunções) (BACELAR DO NASCIMENTO et al., 2005).

Dados os objectivos do projecto, no que respeita à extracção lexical das palavras plenas do *corpus* e dada a falta de possibilidades para fazer uma análise fina das categorias secundárias, optou-se por ter em conta, para a lematização, apenas as etiquetas principais para Verbo, Nome Comum, Adjectivo (não distinguindo, na categoria Verbo, entre verbo principal e verbo auxiliar) e ainda a etiqueta para Estrangeirismos:

Nc	Nome comum
V	Verbo
ADJ	Adjectivo
ESTR	Estrangeirismo

Todo o *corpus* está etiquetado conforme o seguinte exemplo:

Guiné Escrito

```
:E\E\CONJc levaram\LEVAR\Vppi a\A\ARTd memória\MEMÓRIA\Nc ,\-\O
quase\QUASE\ADV toda\TODA\INDv a\A\ARTd memória\MEMÓRIA\Nc .\-\O
Contra\CONTRA\PREP a\A\ARTd razão\RAZÃO\Nc .\-\O E\E\CONJc
quando\QUANDO\ADV das\DE+A\PREP+ARTd cinzas\CINZA\Nc se\SE\CL
resgatou\RESGATAR\Vppi a\A\ARTd esperança\ESPERANÇA\Nc ,\-\O
surgiu\SURGIR\Vppi na\EM+A\PREP+ARTd madrugada\MADRUGADA\Nc
um\UM\ARTi outro\OUTRO\INDv ser\SER\Nc .\-\O Um\UM\ARTi
outro\OUTRO\INDv ser\SER\Nc e\E\CONJc uma\UM\ARTi outra\OUTRA\INDv
vida\VIDA\Nc .\-\O Uma\UM\ARTi vida\VIDA\Nc que\QUE\RELi exigia\EXIGIR\Vii
ser\SER\VB vivida\VIVER\PPA .\-\O Em\EM\PREP plena\PLENO\ADJ
fraternidade\FRATERNIDADE Nc .\-\O Os\O\ARTd cidadãos\CIDADÃO\Nc
que\QUE\RELi disso\-\LPREP+DEMi se\SE\CL deram\DAR\Vppi conta\CONTA\Nc
,\-\O ignorando\IGNORAR\VG as\A\ARTd sequelas\SEQUELA\Nc
da\DE+A\PREP+ARTd pilhagem\PILHAGEM\Nc ,\-\O afirmaram\AFIRMAR\Vppi
vários\VÁRIOS\INDv anos\ANO\Nc mais\MAIS\ADV tarde\TARDE\ADV
que\QUE\RELi foi\IR\Vppi_SER nessa\EM+ESSA\PREP+DEMv
madrugada\MADRUGADA\Nc que\QUE\RELi a\A\ARTd
verdadeira\VERDADEIRO\ADJ vida\VIDA\Nc nasceu\NASCER\Vppi .\-\O
```

Quadro 2 – Excerto do *corpus* África etiquetado

As palavras não reconhecidas pelo anotador – quase todas africanismos que não estavam previstos no léxico do português europeu associado à ferramenta – foram analisadas nos contextos em que ocorriam, anotadas manualmente e posteriormente revistas.

Lematização do *Corpus*

Foram lematizadas todas as formas etiquetadas do *corpus* pertencentes às categorias mencionadas no ponto anterior, constando da lematização a frequência de ocorrência de cada uma. Não foram incluídos na indexação os nomes próprios (antropónimos e topónimos).

Da lematização constam as formas flexionadas antecedidas da cabeça de lema, esta em maiúsculas, e seguidas da categoria gramatical e da respectiva frequência

de ocorrência; na última linha de cada lema, a cabeça de lema em maiúsculas vem seguida da categoria gramatical e da frequência total de ocorrências do lema.

Como é habitual, no caso dos Verbos a cabeça de lema é o infinitivo, no caso dos Nomes e Adjectivos é a forma singular e, naqueles que flexionam em género, a forma do masculino.

O *corpus* África tem, no seu total, 25.359 vocábulos (Nomes, Adjectivos, Verbos e Estrangeirismos) assim repartidos:

- Nomes: 14.466 (57%)
- Adjectivos: 6.268 (25%)
- Verbos: 4.292 (17%)
- Estrangeirismos: 297 (1%)

A repartição destes vocábulos por número de variedades é, percentualmente, a seguinte:

Lemas que ocorreram em: 5 variedades – 25%; 4 variedades – 10%; 3 variedades – 11%; 2 variedades – 15%; 1 variedade – 38%.

Índices contrastivos dos léxicos lematizados com dados de Frequência

Organizaram-se índices contrastivos dos vocábulos lematizados de que se seguem exemplos:

LEMA	FORMA	CAT.	ANG	CV	GUI	MOC	ST	TOTAL
FRUSTRAÇÃO	Frustração	Nc	0	0	1	0	0	1
FRUSTRAÇÃO		Nc	0	0	1	0	0	1
FRUSTRAR	Frustra	V	0	0	1	0	0	1
FRUSTRAR	Frustrada	V	0	0	0	0	2	2
FRUSTRAR	Frustrados	V	0	0	0	0	1	1
FRUSTRAR		V	0	0	1	0	3	4
FRUTA	Fruta	Nc	2	2	0	1	6	11
FRUTA	Frutas	Nc	0	0	0	0	1	1
FRUTA	Frutinha	Nc	0	0	0	0	1	1
FRUTA		Nc	2	2	0	1	8	13
FRUTA-PÃO	fruta-pão	Nc	0	0	0	0	2	2
FRUTAPÃO		Nc	0	0	0	0	2	2

FRUTÍFERO	Frutífera	ADJ	0	1	0	0	0	1
FRUTÍFERO		ADJ	0	1	0	0	0	1
FRUTO	Fruto	Nc	1	1	0	1	3	6
FRUTO	Frutos	Nc	0	0	0	1	1	2
FRUTO		Nc	1	1	0	2	4	8
FUBA	Fubá	Nc	1	0	0	0	0	1
FUBA	Fubás	Nc	2	0	0	0	0	2
FUBA		Nc	3	0	0	0	0	3
FUGA	Fuga	Nc	1	0	0	1	0	2
FUGA		Nc	1	0	0	1	0	2
FUGAR	Fugam	V	1	0	0	0	0	1
FUGAR		V	1	0	0	0	0	1
FUGIR	Foge	V	0	0	1	1	0	2
FUGIR	Fogem	V	0	0	0	1	0	1
FUGIR	Fugia	V	0	0	0	1	0	1
FUGIR	Fugir	V	0	2	0	1	0	3
FUGIR		V	0	2	1	4	0	7
FULANO	Fulana	Nc	1	0	1	0	0	2
FULANO	FulanoNc		2	0	0	4	0	6
FULANO		Nc	3	0	1	4	0	8
FUMAR	Fumam	V	2	0	0	0	0	2
FUMAR	Fumavam	V	1	0	0	0	0	1
FUMAR	Fumo	V	0	0	0	2	0	2
FUMAR		V	3	0	0	2	0	5
FUMAROLA	Fumarolas	Nc	0	5	0	0	0	5
FUMAROLA		Nc	0	5	0	0	0	5
FUNANÁ	Funaná	Nc	0	6	0	0	0	6
FUNANÁ		Nc	0	6	0	0	0	6
FUNÇÃO	Função	Nc	2	2	5	0	6	15
FUNÇÃO	Funções	Nc	0	0	1	0	0	1
FUNÇÃO		Nc	2	2	6	0	6	16

FUNCIONAL	Funcional	ADJ	0	0	0	0	2	2
FUNCIONAL		ADJ	0	0	0	0	2	2
FUNCIONAMENTO	funcionamento	Nc	0	2	0	0	1	3
FUNCIONAMENTO		Nc	0	2	0	0	1	3
FUNCIONAR	Funciona	V	2	1	0	3	3	9
FUNCIONAR	Funcionam	V	1	0	0	0	0	1
FUNCIONAR	Funcionar	V	3	8	4	0	5	20
FUNCIONAR	Funcionava	V	0	1	1	0	0	2
FUNCIONAR	Funciono	V	0	0	0	0	2	2
FUNCIONAR	Funcionou	V	0	0	0	0	1	1
FUNCIONAR		V	6	10	5	3	11	35

Quadro 3 – Quadro contrastivo da lematização do *corpus* África (oral) com dados de frequência e repartição por variedade (excerto)

LEMA	FORMA	CAT.	ANG	CV	GUI	MOC	ST	TOTAL
MANGA	Mangas	Nc	8	5	4	8	1	26
MANGA	manguinha	Nc	0	0	0	0	3	3
MANGA	manguinhas	Nc	0	0	0	0	2	2
MANGA		Nc	20	8	6	14	16	64
MANGAL	Mangais	Nc	1	0	0	1	1	3
MANGAL	Mangal	Nc	3	0	1	4	0	8
MANGAL		Nc	4	0	1	5	1	11
MANGANÊS	Manganês	Nc	3	0	0	0	0	3
MANGANÊS		Nc	3	0	0	0	0	3
MANGAR	Mangar	V	2	0	0	0	0	2
MANGAR	mangarmos	V	1	0	0	0	0	1
MANGAR		V	3	0	0	0	0	3
MANGEDOURA	mangedoura	Nc	0	1	0	0	0	1
MANGEDOURA		Nc	0	1	0	0	0	1
MANGERONA	Mangerona	Nc	0	1	0	0	0	1
MANGERONA		Nc	0	1	0	0	0	1
MANGIAGO	Mangiago	ADJ	0	0	1	0	0	1
MANGIAGO		ADJ	0	0	1	0	0	1

MANGO	Mango	Nc	0	0	19	0	0	19
MANGO	Mangos	Nc	0	0	2	0	0	2
MANGO		Nc	0	0	21	0	0	21
MANGONHEIRO	mangonheiro	Nc	0	0	0	0	1	1
MANGONHEIRO		Nc	0	0	0	0	1	1
MANGOSTÃO	Mangostão	Nc	0	0	0	0	1	1
MANGOSTÃO		Nc	0	0	0	0	1	1
MANGUAVAVA	manguavavas	Nc	0	0	0	1	0	1
MANGUAVAVA		Nc	0	0	0	1	0	1
MANGUÇO	Manguços	Nc	0	0	0	1	0	1
MANGUÇO		Nc	0	0	0	1	0	1
MANGUEIRA	Mangueira	Nc	2	0	2	5	6	15
MANGUEIRA	mangueiras	Nc	2	1	1	6	6	16
MANGUEIRA		Nc	4	1	3	11	12	31
MANGUEIRO	Mangueiro	Nc	0	0	4	0	0	4
MANGUEIRO		Nc	0	0	4	0	0	4

Quadro 4 – Quadro contrastivo da lematização do *corpus* África oral e escrito com dados de frequência e repartição por variedade (excerto)

LEMA	CAT.	ANG	CV	GUI	MOC	ST	TOTAL
FRUIR	V	1	1	0	0	0	2
FRUSTRAÇÃO	Nc	4	14	16	5	2	41
FRUSTRADO	ADJ	1	0	0	4	0	5
FRUSTRANTE	ADJ	4	2	0	0	1	7
FRUSTRAR	V	10	5	16	8	7	46
FRUTA	Nc	31	3	14	18	18	84
FRUTA-PÃO	Nc	0	0	0	0	2	2
FRUTARIA	Nc	1	0	0	0	0	1
FRUTEIRA	Nc	2	0	0	2	0	4
FRUTÍCOLA	ADJ	0	0	0	4	0	4
FRUTICULTOR	Nc	0	0	0	4	0	4
FRUTICULTURA	Nc	0	0	0	3	0	3
FRUTÍFERO	ADJ	1	1	1	0	2	5
FRUTIFICAR	V	0	2	0	0	0	2
FRUTÍVORO	ADJ	1	0	0	0	0	1

FRUTO	Nc	46	30	29	33	53	191
FRUTUOSO	ADJ	0	2	3	0	0	5
FUBA	Nc	9	0	0	0	1	10
FUÇA	Nc	2	0	0	1	0	3
FUEL	ESTR	0	0	0	1	0	1
FUGA	Nc	22	25	19	31	30	127
FUGACIDADE	Nc	0	1	0	1	0	2
FUGAR	V	1	0	0	0	0	1
FUGAZ	ADJ	2	2	1	2	2	9
FUGIDIO	ADJ	2	1	0	3	2	8
FUGIR	V	57	75	44	96	28	300
FUGITIVO	ADJ	0	0	1	2	1	4
FUGITIVO	Nc	0	2	1	2	1	6
FUJÃO	Nc	0	0	0	0	2	2
FULANIZADO	ADJ	0	0	0	1	0	1
FULANO	Nc	5	19	12	8	3	47
FULCRAL	ADJ	4	0	3	0	0	7
FULCRO	Nc	0	5	1	0	0	6
FULGIR	V	0	0	0	1	0	1
FULGOR	Nc	2	2	2	1	2	9
FULGURÂNCIA	Nc	1	0	0	0	0	1
FULGURANTE	ADJ	2	0	1	1	0	4
FULIGEM	Nc	0	2	0	0	1	3
FULIGINOSO	ADJ	0	0	0	1	0	1
FULMINANTE	ADJ	0	0	3	0	3	6
FULMINAR	V	0	1	1	6	1	9
FUMAÇA	Nc	0	4	1	0	1	6
FUMADOR	ADJ	0	0	0	1	1	2
FUMADOR	Nc	2	0	0	2	2	6
FUMAGEM	Nc	0	0	0	0	7	7
FUMANTES	Nc	0	0	0	18	0	18
FUMAR	V	16	19	1	21	10	67
FUMARAÇA	Nc	1	0	0	0	0	1
FUMARADA	Nc	0	1	0	0	0	1
FUMARENTO	ADJ	1	0	0	0	1	2
FUMAROLA	Nc	0	9	0	0	0	9
FÚMBUA	Nc	3	0	0	0	0	3
FUMEGANTE	ADJ	1	1	0	0	5	7
FUMEGAR	V	0	0	0	0	2	2
FUMO	Nc	12	8	4	25	15	64
FUNANÁ	Nc	0	6	0	0	0	6

FUNANTE	Nc	1	0	0	0	0	1
FUNÇÃO	Nc	205	318	265	120	288	1196
FUNCHE	Nc	0	1	0	0	0	1
FUNCHO	Nc	0	0	0	5	0	5
FUNCIONAL	ADJ	24	23	6	9	3	65
FUNCIONALIDADE	Nc	4	14	1	1	2	22
FUNCIONALISMO	Nc	0	9	1	0	5	15
FUNCIONALIZAR	V	0	1	0	0	0	1
FUNCIONAMENTO	Nc	79	136	66	79	241	601
FUNCIONAR	V	152	105	87	61	67	472

Quadro 5 – Quadro contrastivo do vocabulário (cabeças de lema) do *corpus* África oral e escrito com dados de frequência e repartição por variedade (excerto)

Concordâncias

Um dos resultados mais importantes deste trabalho é a disponibilização de concordâncias, para consulta, de todas as ocorrências dos cinco *corpora*, subdivididas pelas cinco variedades e pelos registos oral e escrito.

mos se essas teorias também... de	**facto**	corresponde ou não à verdade.
nto quer comportar assim mas pelo	**facto**	de ele ter de ele saber que ho
atirado com pedra no liceu, pelo	**facto**	de eu o ter... sancionado com
não sei se seja, se isso pode ser	**facto**	de eu habituar mais com aquilo
de S. Tomé para o Portugal, é um	**facto**	de, quando em São Tomé dizer-s
m o próximo paciente mas sim pelo	**facto**	de saber como eu já tinha dito
de direito. mas não, não foi pelo	**facto**	de ter me aconselhado que me i
eh, por esse motivo eu acho que o	**facto**	de ter conseguido para mim é u
certo modo abstracto, porque pelo	**facto**,	eh, embora, embora eu tenha c
..] numa estação de comboio pelo	**facto**	eh... pelo facto eu não consig
ele tem que ter conhecimento dos	**factos**.	ele tem que saber, por exemp
de comboio pelo facto eh... pelo	**facto**	eu não consigo explicar as raz
ativo, bom! no aspecto positivo o	**facto**	importante é eu ter conseguido
a fazer o curso de medicina. e, o	**facto**	não. eu acho que, eh são, são
ver com a ligação que ele faz dos	**factos**	que acontecem dia a dia nos p
pessoalmente não tenho. não tenho	**factos**	que me marcaram tanto no aspe
a-se mal não sei que mas elas são	**factos**	que nos a sociedade devia por
ta, uma cor bonita para regar ou	**para**	pintar o seu quarto. imaginação, cr
os fazer? um molhinho para regar,	**para**	pintar, para dar o toque final ao n
esbravamento. [...] isso sim, mas	**para**	plantação mesmo é à mão. e nós não

ensino que os professores não tem	**para**	poderem transmitir convenientemente
umas peças. tem um macaco próprio	**para**	poder tocar, esticar carro, alisar
mo a pessoa que vai, por exemplo,	**para**	Ponta Firme não encontra-se cobra.
ubon-tunhá' . é um tubarão que dá	**para**	qualquer comida, esse 'tubon-tunhá'
bém meter a música deles em acção	**para**	que a nossa população comece a dist
a que a prá [...] , a prá [...] ,	**para**	que a prática continua. é assim: nó
re contacto, dialogando, falando,	**para**	que a prá [...] , a prá [...] , par
Âmbito profissional contribuírem	**para**	que a pessoa possa ter ter uma vida
nco na altura impunha, quer dizer	**para**	que as pessoas não comessem determi
um poucochinho de água do feijão,	**para**	quê? já vão ver. a nossa sertã está
s, cores, misturar este bocadinho	**para**	quê? para dar o toque final e engro
es no, no serviço activo não é, e	**para**	quê, para aqueles que trabalham con
Alizado. mas as condições de base	**para**	que essa rea [...] , essa realizaçã
didácticos mínimos indespensáveis	**para**	que esses meninos possam estar na e
u conseguir uma formação de base,	**para**	que eu venha a conseguir-me, eh, re

Quadro 6 – Excerto de concordâncias das formas dos vocábulos *facto* e *para* extraídas do *corpus* oral de São Tomé

não te procuro muito para não te	**pegar**	a má disposição, tens muito
as fica muito caro. - Não sei se	**pegava**	bem - disse Vítor, rompendo
as olha só: esse Zé Maria quando	**pega**	da viola, parece até chora, n
o respeito. Vá lá dizer-lhe para	**pegar**	depressa nas imbambas dele e
ede são o próprio início, podes	**pegar**	de qualquer lado, então eu n
o. "Espero que este truque não	**pegue.**	Descemos e mais nada. Não v
bém ninguém pegar dum lado e só	**pegarem**	do outro, o lado que não p
s novas porque se também ninguém	**pegar**	dum lado e só pegarem do out
s moradores considerados aptos a	**pegarem**	em armas. Em 1574 a cidade
a contar o filho jeovou para não	**pegar**	em armas e isso até tem um n
cima de duas dezenas de pessoas.	**Peguemos**	em duas pessoas desse gru
, nessas idas dele, que o Adolfo	**pegou**	então a doença que lhe foi m
Pegarem do outro, o lado que não	**pegou**	está embora lixado, não é? a
nteiro. Vi matarem um ladrão. E	**pegaram-lhe**	fogo. O bispo fez um a
de Dominus. Não é por aí que te	**pegam.**	Mas, claro, podiam começar
eja, como eles queriam. Mussunda	**pegara**	Miguel num braço e, em baix
sorria no sol. Era sempre assim:	**pegava**	miúdo Zito na mão, qualquer
ta. E o Polobochi nem sequer lhe	**pegou**	na barriga para uma leve umb
ar de, e com o inglês que tem, e	**pegando**	na coisa à sério, Hendrick
o no gesto da mão fez questão de	**pegar**	na mão do albino e pôr lá na
tando depois um colégio de madres	**em**	Sá da Bandeira (actual Lubango) , ond
tado, ninguém se preocupava muito	**em**	saber como os acidentes ocorriam. Mal

" . Nem vale a pena preocupar-nos	**em**	saber se todas as pessoas são capazes
iro projecto do governo) consiste	**em**	saber se as terras periféricas já apr
os que tavam só preocupados masé	**em**	sair dali, fosse a KotaDasAbelhas ou
acima de tudo, disponibilizando-o	**em**	sala de leitura. O Empréstimo Domicil
da, idosas, a serem transformadas	**em**	salões de beleza, pedicure e manicure
eguiu. O garoto ficou a esvair-se	**em**	sangue no chão. A bicicleta havia sid
bas paralisadas antes de 1992, e	**em**	Sanza Pombo uma fábrica de descasque

Quadro 7 – Excerto de concordâncias das formas dos vocábulos *pegar* e *em* extraídas do *corpus* escrito de Angola

Conclusão

Os resultados deste trabalho são complementares: por um lado, a indexação de palavras isoladas e, por outro lado, as concordâncias, mediante as quais, dada uma forma lexical ou um lema, se podem identificar os seus contextos de uso nos *corpora* comparáveis. Com base nas concordâncias podem analisar-se as palavras em função das construções em que ocorrem, dos valores semânticos que apresentam, dos grupos de coocorrentes em que se inserem, dos registos de língua em que são utilizados.

Este é um primeiro contributo para a descrição linguística das cinco variedades africanas do português e para a construção de léxicos e de gramáticas individuais ou contrastivas. Para além das especificidades linguísticas, o estudo dos *corpora* permite ainda a observação de aspectos culturais ligados ao uso dos africanismos.

Agradecimentos

Agradeço a colaboração neste artigo a toda a equipa do projecto e muito particularmente às colegas Luísa Alice Santos Pereira e Antónia Estrela.

BACELAR DO NASCIMENTO, M. F. Comparable corpora and lexical variation in the African varieties of Portuguese. *Alfa*, São Paulo, v.50, n.2, p.189-204, 2006.

- *ABSTRACT: This paper presents the results of a project focused on the compilation of* corpora *of the five Portuguese African varieties (Angola, Cape Verde, Guinea-Bissau, Mozambique and São Tomé and Príncipe) and on the extraction of* corpus*-driven lexicons for each of these varieties. The five* corpora *(total of 3.200.124 words) are comparable in size, composition and chronology. This project provides an answer to the important need of Linguistic resources (*Corpora *and Lexicons) for Portuguese. In fact, a high number of large* corpora *and lexicons*

were already available for European and Brazilian Portuguese varieties, while such resources were almost non-existent for African varieties (Mozambique is an exception). Those five comparable corpora *will now enable objective linguistic analysis of each variety, as well as contrastive studies between these African varieties, and also between those and the European and Brazilian ones. The project, named Linguistic Resources for the Study of Portuguese African Varieties, was undertaken by the* Corpus *Linguistics Group at the Centro de Lingüística da Universidade de Lisboa (CLUL) and by a team from the Centro de Física Teórica e Computacional (Universidade de Lisboa), with consultancy work by Perpétua Gonçalves, of the Universidade Eduardo Mondlane (Mozambique).*

- *KEYWORDS:* Corpus *Africa; lexicons; variation; Portuguese.*

Referências bibliográficas

BACELAR DO NASCIMENTO, M. F. et al. The Portuguese *corpus*. In: CRESTI, E.; MONEGLIA, M. (Ed.) *C-ORAL-ROM*: Integrated reference corpora for spoken romance languages. Amsterdam: John Benjamins, 2005. (Studies in Corpus Linguistics, 15).

BIBER, D. et al. *Longman grammar of spoken and written English*. London: Longman, 1999.

PEYAWARY, A. S. *The core vocabulary of international English*: a *corpus* approach. Bergen: The Humanities Information Technology Research Programme, 1999.

REPPEN, R.; FITZMAURICE, S. M.; BIBER, D. (Ed.) *Using* corpora *to explore linguistic variation*. Amsterdam: John Benjamins, 2002.

Bibliografia consultada

BACELAR DO NASCIMENTO, M. F. et al. The African varieties of Portuguese: compiling comparable corpora and analyzing data-derived lexicon. In: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 5., 2006, Genova. *Proceedings*... Génova: LREC, 2006. p.1791-94.

BACELAR DO NASCIMENTO, M. F. et al. As variedades africanas do português: um corpus comparável. In: *SIMPOSIO INTERNACIONAL DE COMUNICACIÓN SOCIAL*, 10., 2006, Santiago de Cuba. *Actas*... Santiago de Cuba: Ministerio de Ciencia, Tecnología y Medio Ambiente, 2007. p.232-37.

BIDERMAN, M. T. C. O português brasileiro e o português europeu: identidade e contrastes. *Revue Belge de Philologie et d'Histoire*, Bruxelles, v.79, p.963-975, 2001.

GONÇALVES, P. Panorama geral do português de Moçambique. *Revue Belge de Philologie et d'Histoire*, Bruxelles, v.79, p. 977-990, 2001.

GONÇALVES, P.; STROUD, C. (Org.) *Panorama do português oral de Maputo*: vocabulário básico do português (espaço, tempo e quantidade): contextos e prática Pedagógica. Maputo: INDE, 2000. v.4.

GREENBAUM, S. The international *corpus* of English. *ICAME Journal*, Bergen, v.14, p.106-108, 1990.

HOFLAND, K.; JOHANSSON, S. *Word Frequencies in British and American English*. Bergen: Norwegian Computing Centre for Humanities, 1982.

LEITNER, G. The Kolhapur *corpus* of Indian English: intravarietal description and/or intervarietal comparison. In: JOHANSSON, S.; STENSTRÖM, A-B. (Ed.) *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter, 1991.

MENDES, A.; BACELAR DO NASCIMENTO, M. F.; PEREIRA L. As concordâncias: um instrumento para a aprendizagem da língua portuguesa a partir de dados de um *corpus*. *Idiomático*: Revista digital de didáctica de PLNM, Lisboa, Centro Virtual Camões, Instituto Camões, v.1, 2004.

MENDES, A.; AMARO, R., BACELAR DO NASCIMENTO, M. F. Morphological tagging of a spoken Portuguese *corpus* using available resources. In: BRANCO, A.; MENDES, A.; RIBEIRO, R. (Ed.) *Language technology for Portuguese*: shallow processing tools and resources. Lisboa: Colibri, 2004.

MOTA, M. A.; BACELAR DO NASCIMENTO, M. F. Le portugais dans ses variétés. *Revue Belge de Philologie et d'Histoire*, Bruxelles, v.79, p.931-952, 2001.

STROUD, C.; GONÇALVES, P. (Org.) *Panorama do português oral de Maputo*: a construção de um banco de "erros". Maputo: INDE, 1997. v.2.