MARIA FERNANDA BACELAR DO NASCIMENTO*
JOSÉ BETTENCOURT GONÇALVES*
LUÍSA PEREIRA*
ANTÓNIA ESTRELA*
AFONSO PEREIRA*
RUI SANTOS**
SANCHO M. OLIVEIRA**
Consultora: PERPÉTUA GONÇALVES***
Centro de Linguística da Universidade de Lisboa – CLUL, Portugal*
{fbacelar.nascimento, jose.bettencourt, luisa.alice, antonia.estrela, afonso.pereira }@clul.ul.pt
Centro de Física Teórica e Computacional, Universidade de Lisboa, Portugal**
{rsantos,smo}@cii.fc.ul.pt
peta@zebra.uem.mz***

As variedades africanas do português: um corpus comparável

Introdução

Ocupando o Português, sensivelmente, o quinto lugar entre as línguas mais faladas do mundo, repartido pelos cinco continentes, por oito estados independentes, quer como língua nacional, quer como língua oficial, é natural que apresente considerável nível de variação. O recurso a *corpora* há muito que é reconhecido como uma valiosa fonte de comparação na descrição das variedades de uma língua. A título de exemplo, veja-se o "LOB Corpus" e o "Brown Corpus", o primeiro do inglês britânico, constituído em Lancaster (G. Leech), Oslo (S. Johanson) e Bergen (K. Hofland), o segundo do inglês americano, constituído por W. N. Francis e H. Kucera, na Brown University, Providence, ambos com cerca de um milhão de palavras, recolhidos no mesmo ano, 1961, e contendo géneros e dimensão de amostragens idênticos. O "Kolhapur Indian Corpus" é um corpus do inglês indiano, constituído nos mesmos moldes dos dois anteriores, excepto no que se refere à data: 1978. Estes três "corpora comparáveis" têm sido o recurso para identificar o vocabulário básico do inglês internacional, o que é considerado de fundamental importância para o desenvolvimento de materiais usados no ensino da língua inglesa (ELT)[1]. Vários outros casos poderiam ser evocados. De entre eles, salientaremos "The International Corpus of English" (ICE), construído por treze grupos nacionais (entre eles, Austrália, Canadá, África Oriental, Índia, Jamaica, Nova Zelândia, Nigéria, Filipinas, Reino Unido, Estados Unidos da América) sob a coordenação de S. Greenbaum, University College, Londres. Mais próximo de nós, merece referência o "Corpus de Referencia del Español Actual" (CREA), que visa produzir recursos para estudos comparativos das linguagens usadas nos países de expressão espanhola.

Estes corpora, tal como outros não citados, encontram-se disponíveis para servirem de base a descrições e teorização linguísticas e reflectem o papel dos dados naturais no que toca à investigação e ao desenvolvimento de pedagogias da língua com base na linguística de corpus.

Considerando o panorama dos países lusófonos, nota-se uma notável disparidade na área dos Recursos Linguísticos, no que se refere à publicação de estudos sobre as variedades do Português: existem numerosos estudos comparativos abrangendo o português do Brasil e o português europeu, sendo quase total a ausência dos estudos de cada uma das variedades africanas[2] e de estudos comparativos relacionados. O projecto agora em desenvolvimento no Centro de Linguística da Universidade de Lisboa – CLUL[3] visa contribuir para que se atenue esse desfasamento, fornecendo recursos linguísticos que permitam descrever com objectividade as cinco variedades africanas do português.

Esta comunicação destina-se a apresentar os trabalhos deste projecto.

O corpus - sua constituição

O Centro de Linguística da Universidade de Lisboa - CLUL tem disponível um corpus: o "Corpus de Referência do Português Contemporâneo" (CRPC), actualmente com a dimensão de cerca de 334.500.000 de palavras[4]. Neste corpus estão representadas variedades do português de Portugal, Brasil, Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe, assim como Timor, Macau e Goa, sendo o subcorpus escrito de 332 milhões de palavras e o subcorpus oral de 2,5 milhões. Quanto aos géneros representados no subcorpus escrito, estes vão do Livro a Correspondência, Panfletos e Miscelânea, passando por Jornais e Revistas, Jornal das Sessões Parlamentares e Acórdãos do Supremo Tribunal de Justiça, incluindo ficção, discurso técnico-científico, didáctico e discurso corrente. O subcorpus oral contém, por sua vez, discurso formal e discurso informal.

O presente trabalho baseia-se neste corpus geral, o CRPC, que é um corpus monitor e visa a constituição, tratamento, análise e disponibilização *on-line* de um corpus das variedades africanas do português, constituído por cinco subcorpora comparáveis, correspondentes às variedades dos cinco países designados por PALOP

1 Peyaway, 1999.
2 Exceptua-se o português de Moçambique (Cf. Stroud, C. and Gonçalves, P. (org.) (1997); Gonçalves, P. & Stroud, C. (org.), (1999, 2000).
3 Projecto "Recursos linguísticos para o estudo das variedades africanas do português" (PLUS/LIN/50270/2003), financiado Pela Fundação para a Ciência e a Tecnologia e pela Fundação Calouste Gulbenkian.
4 Cf. Quadro 1.

(Países Africanos de Língua Oficial Portuguesa): Angola, Cabo Verde, Guiné-Bissau, Moçambique e São Tomé e Príncipe. Os materiais extraídos deste corpus permitirão a investigadores, professores, autores de materiais (dicionários, gramáticas, manuais) e alunos a acessibilidade a dados autênticos que possibilitarão que, pela primeira vez, se façam estudos descritivos empíricos de cada uma das cinco variedades mencionadas. Para além do que acaba de ser dito, estes materiais proporcionarão também estudos comparativos intra e intercorpora das variedaddes do português que permitirão visualizar, por um lado, a variação e o resultado de diferenças discursivas e pragmáticas e, por outro lado, aspectos de unidade linguística do português nas suas cinco variedades africanas. Os cinco subcorpora estabelecidos para este estudo são comparáveis em dimensão (600.000 palavras cada) e cronologicamente (últimos 30 anos) assim como em tipo e género[5].

Do que acaba de ser exposto se pode ver que o principal objectivo do nosso trabalho é a compilação de uma base de conhecimento disponível para todos aqueles que necessitem de estudos comparativos relativos às cinco variedades africanas do português. Com os resultados deste projecto, tornar-se-á possível o acesso a recursos organizados a partir de dados empíricos, assim como a estudos descritivos do português africano. O trabalho permitirá ainda uma visão global do panorama da língua portuguesa usada em contextos multilingues, como língua materna, língua segunda ou língua estrangeira.

Como atrás foi dito, o CRPC é um corpus monitor e contém amostragens das variedades dos cinco países africanos de expressão portuguesa. Veja-se no Quadro 1 a dimensão de cada um dos subcorpora das variedades do português no CRPC:

CRPC	
Variedades do Português	Dimensão de cada subcorpus
Angola	c. 15.800.000
Cabo Verde	c. 3.200.000
Guiné-Bissau	c. 800.000
Moçambique	c. 2.900.000
São Tomé e Príncipe	c. 900.000
Brasil	c. 23 000 000
Portugal	c. 280 000 000
Goa	c. 2 000
Macau	c. 7 800 000
Timor	c. 140 000

Quadro 1: Constituição do CRPC

Contudo, estas amostragens não são equilibradas no que concerne à sua dimensão e estrutura. Para atingir tal equilíbrio no corpus organizado para o presente projecto, indispensável para estudos comparativos, apenas foi possível chegar às 600.000 palavras para cada uma das variedades africanas. Um factor a considerar foi o projecto ter tido a curta duração de dois anos. Assim, foram reutilizados, sempre que possível, os materiais já existentes; mas, tendo em mente o equilíbrio da estrutura dos subcorpora no respeitante a dimensão, domínios semânticos, género de texto e período de tempo, foi necessário proceder à recolha de novos materiais tanto escritos como orais. O desenho de cada um dos cinco subcorpora é o que a seguir se apresenta:

Tipo e género de texto		**Percentages/ Nº de Palavras**
ESCRITO	Livro	20% (120.000)
	Jornais/Revistas	50% (300.000)
	Varia	26% (156.000)
ORAL	Formal/informal	4% (24.000)
TOTAL		**100% (600.000)**

Quadro 2: Constituição do corpus de cada uma das cinco variedades africanas do português

Lematização das formas

Uma vez constituído o corpus, uma lista de formas alfabeticamente ordenadas é constituída para cada um dos subcorpora e, subsequentemente, é feita a lematização automática das formas. Segue-se uma revisão manual desta lematização automática. Por este processo constituem-se cinco listas de vocabulário correspondentes aos cinco países cujas variedades do português são objecto deste estudo, podendo assim constituir-se a base de dados que permitirá o tratamento estatístico dos resultados, indexação de cada palavra do corpus ao seu contexto, permitindo assim uma extracção de concordâncias bastante operativa. Nesta base de dados é possível obter alguma informação respeitante ao texto tal como autor, editor, data, no caso dos livros, ou, simplesmente, a fonte no caso de outros materiais.

Apresenta-se, a seguir o Quadro 3, onde se podem ver simultaneamente diversas situações:
- um verbo que apenas existe nas variedades apresentadas;
- o modelo de lematização;
- a comparação entre duas variedades[6]:

5 Cf. Quadro 2.
6 Este tipo de comparação pode ser usado para qualquer número de variedades.

		MOÇAMBIQUE			ANGOLA		
DESCONSEGUIR	Categ.	Escrito	Oral	Total	Escrito	Oral	Total
desconsegue	V	0	0	0	3	0	3
desconseguem	V	1	0	1	0	0	0
desconsegues	V	0	0	0	1	0	1
desconseguia	V	0	0	0	2	0	2
desconseguir	V	2	0	2	1	1	2
desconseguiram	V	0	0	0	1	0	1
desconseguiu	V	0	0	0	2	0	2

Quadro 3: Verbo "Desconseguir", comparado entre duas variedades

Resultados

Os resultados deste projecto estão disponíveis na página *web* do CLUL (www.clul.ul.pt) para pesquisa *on-line*, incluindo documentação relativa a objectivos, metodologias e aos próprios resultados.

Estão, assim, disponíveis *on-line* os seguintes materiais:

1. Concordâncias em formato KWIC de todas as palavras do corpus, organizadas por subcorpora de cada variedade;
2. Indexação de todas as palavras (lema e respectivas formas) que ocorrem no corpus, com dados de frequência e de distribuição por subcorpora;
3. Descrição comparativa do vocabulário dos diversos subcorpora, resultado de alguns estudos quantitativos e estatísticos.

Contrastivamente com a norma europeia do português, nota-se, nas variedades africanas, não só diferenças consideráveis, mas também regularidades. Muitas vezes ocorre a forma de acusativo onde se esperaria a do dativo e vice-versa, como se pode ver no Quadro 4, ou a diferença de regência verbal, exemplificada no Quadro 5:

res, e um monte de abelhas a	**lhe**	bicarem as roupas, ele parado, se tu
ivo, que a sala, em uníssono,	**lhe**	brindava. Do alto do seu metro e
ramelados que você no olhar	**lhe**	busca e encontra a ternura - aí você
brincadeira, quem ia pinar pra	**lhe**	buscar? Sabes? AdolfoDido. Aquilo
"ti Raúl". Com toda a justeza	**lhe**	caberia o título de capital da cultura

Quadro 4: Construção com o clítico lhe em discurso escrito de Angola

o seu perseguidor amoroso. Este	pegou-lhe	**na** mão. Entre a exigênc
são copiadas à máquina. Menina Maria	pega	**da**
—Você assente-se, senhor Xalino. Ela	pega	**de** um pano, passa-o sobre o
corpo talhado de vergônteas. Mamãe	pegou-me	com uma indignação que
de pau. Aquele ali serve. Zé-do-colete	pega	**do** pau e começa a bater no m

Quadro 5: Regência do verbo pegar em discurso oral da Guiné-Bissau

Foram observados exemplos da diferença das variedades africanas relativamente ao português europeu.

Apresentam-se a seguir os Quadros 6, 7 e 8, onde se podem observar exemplos da ocorrência dos 20 verbos, substantivos e adjectivos mais frequentes nos subcorpora deste projecto, com as características acima descritas, das variedades de Angola, Moçambique e Cabo Verde[7]:

Classe:	Lema:	Angola	Moçambique	Cabo Verde
Verbos	FREQUÊNCIA DECR	**Ser**	**ser**	**Ser**
		ter	**ter**	**Ter**
		estar	**estar**	**Estar**
		poder	**ir**	**Dizer**
		dizer	**dizer**	**Fazer**
		ir	**fazer**	**Poder**
		fazer	**poder**	**Ir**
		haver	**haver**	**Haver**
		dever	**dever**	**Saber**
		saber	**querer**	**Dar**
		falar	**saber**	**Querer**
		querer	**ver**	**Ver**
		dar	**dar**	**Dever**

7 Destacam-se em negrito os lemas comuns às três variedades.

	E S C E N T E	**ver**	**vir**	**Vir**
		vir	**ficar**	**Ficar**
		ficar	**passar**	**Falar**
		passar	**falar**	**Passar**
		deixar	**chegar**	**Deixar**
		chegar	continuar	**Chegar**
		encontrar	**deixar**	Pensar

Quadro 6: 20 Verbos mais frequentes

Classe:	Lema:	Angola	Moçambique	Cabo Verde
Nomes	F R E Q U Ê N C I A D E C R E S C E N T E	**ano**	**país**	**País**
		país	**ano**	**Ano**
		dia	**dia**	**Dia**
		direito	mulher	**Vez**
		vez	**vez**	Coisa
		tempo	**homem**	**Tempo**
		pessoa	desenvolvimento	**Pessoa**
		província	**pessoa**	**Vida**
		caso	**parte**	**Homem**
		entrevista	informação	Forma
		poder	**tempo**	Política
		parte	**vida**	**Casa**
		coisa	**casa**	Partido
		vida	projecto	Desenvolvimento
		casa	processo	Facto
		política	**trabalho**	Gente
		forma	filho	**Parte**
		homem	**caso**	**Trabalho**
		trabalho	cidade	Poder
		população	sector	**Caso**

Quadro 7: 20 Substantivos mais frequentes

Classe:	Lema:	Angola	Moçambique	Cabo Verde
Adj.	F R E Q U Ê N C I A D E C R E S C E N T E	angolano	**novo**	**Grande**
		novo	moçambicano	**Novo**
		grande	**nacional**	**Social**
		nacional	**grande**	**Nacional**
		próprio	**último**	**Bom**
		maior	**maior**	Melhor
		social	**bom**	**Económico**
		bom	melhor	**Maior**
		último	**social**	**Próprio**
		melhor	**próprio**	**Político**
		público	geral	cabo-verdiano
		económico	eleitoral	Certo
		principal	internacional	**Último**
		importante	**público**	Importante
		político	**velho**	**Público**
		africano	branco	Necessário
		internacional	**político**	**Velho**
		geral	antigo	Cultural
		velho	próximo	Português
		português	**económico**	Possível

Quadro 8: 20 Adjectivos mais frequentes

É interessante notar que, no que respeita à lista dos 20 verbos mais frequentes em cada uma das três variedades exemplificadas, 19 lemas são comuns às três variedades, diferindo apenas na posição que ocupam na respectiva ordem de frequência. No caso dos substantivos e adjectivos o número de lemas comuns incluídos entre os 20 mais frequentes é apenas de 12, continuando a verificar-se a não coincidência quanto ao lugar que ocupam na respectiva ordem de frequência.

No Quadro 9, que a seguir se apresenta, faz-se o mesmo tipo de comparação, apenas para os lemas verbais, mas incluindo as variedades do português europeu e do português do Brasil:

País	Portugal		Brasil	Angola	Moçambique	Cabo Verde
Dimensão e tipo de discurso	300.000 oral	16.000.000 escrito	700.000 oral	600.000 oral e escrito	600.000 oral e escrito	600.000 oral e escrito
	ser	**ser**	**ir**	**ser**	**ser**	**Ser**
	ter	**ter**	**dizer**	**ter**	**ter**	**Ter**
	estar	**estar**	**ter**	**estar**	**estar**	**Estar**
	ir	**fazer**	**fazer**	**poder**	**ir**	**Dizer**
	haver	**ir**	**ser**	**dizer**	**dizer**	**Fazer**
	fazer	**haver**	achar	**ir**	**fazer**	**Poder**
	dizer	**poder**	**estar**	**fazer**	**poder**	**Ir**
	poder	**dizer**	saber	**haver**	**haver**	**Haver**
	saber	**dar**	**poder**	dever	dever	**Saber**
	achar	**ver**	**dar**	**saber**	querer	**Dar**
	ver	dever	**ficar**	falar	**saber**	Querer
	dar	**saber**	falar	querer	**ver**	**Ver**
	querer	querer	**ver**	**dar**	**dar**	Dever
	vir	ficar	**haver**	**ver**	**vir**	**Vir**
	ficar	**vir**	trabalhat	**vir**	**ficar**	**Ficar**
	falar	passar	vir	**ficar**	passar	Falar
	começar	deixar	entender	passar	falar	Passar
	passar	formar	chegar	deixar	chegar	Deixar
	pensar	encontrar	sair	chegar	continuar	Chegar
	gostar	chegar	gostar	encontrar	deixar	Pensar

Quadro 9: 20 Verbos mais frequentes – comparação alargada

A observação deste Quadro 9 sugere que apesar de especificidades de cada uma das variedades em análise, há um fundo comum tendencialmente estável, que permite captar o movimento entre a unidade e a diversidade entre uma língua e as suas variedades.

Trabalhos futuros

Além dos resultados referidos disponíveis *on-line* a partir de Dezembro de 2006, é intenção da equipa do projecto debruçar-se futuramente sobre os seguintes trabalhos gramaticais:

- **Estudo contrastivo sobre a colocação e selecção de clíticos**

Contrastivamente com a norma europeia do português, notam-se, nas variedades africanas, diferenças consideráveis no que respeita a selecção de formas de objecto directo e indirecto. Muitas vezes ocorre a forma de acusativo onde se esperaria a do dativo e vice-versa.

- **Perda de formas casuais**

Estudo do processo de perda de formas casuais pronominais, de um ponto de vista contrastivo. Por um lado entre as diferentes variedades do português e por outro lado entre os discursos oral e escrito.

- **Regências verbais nas variedades do português**

As consideráveis diferenças de regência no português do Brasil, no português europeu e nas variedades africanas são um facto que se constata à vista desarmada; só o recurso a um corpus de dados autênticos permitirá mostrar, com objectividade, os diferentes fenómenos: presença/ausência de preposição e casos de diferença de preposição.

- **Processos derivacionais no português africano**

O objectivo desta pesquisa será, por um lado, a comparação, através do estudo de processos regulares de derivação no português, de novas formas lexicais criadas nas cinco variedades e, por outro lado, a comparação dos processos derivacionais usados.

- **Estudo contrastivo de lexias complexas**

Com base no projecto COMBINA-PT[8] será feita a análise de diferentes tipos de associações de palavras, do ponto de vista do grau de relações colocacionais estabelecidas pelo uso, desde a fixidez absoluta (idiomatismos) até à coocorrência preferencial.

8 Projecto Combinatórias Lexicais do português (COMBINA-PT) - POCTI/LIN/48465/2002, finnaciado pela Fundação para a Ciência e a Tecnologia, em desenvolvimento no CLUL, disponível em: www.clul.ul.pt.

Coordenação do projecto:

Maria Fernanda Bacelar do Nascimento

Financiamento

O Projecto "Recursos Linguísticos para o Estudo das Variedades Africanas do Português" (PLUS/LIN/50270/2003) é financiado Pela Fundação para a Ciência e a Tecnologia e pela Fundação Calouste Gulbenkian.

Agradecimentos

• CIDAC – Centro de Informação e Documentação Amílcar Cabral.
• Departamento de Literaturas Africanas da Faculdade de Letras da Universidade de Lisboa.
• Embaixada de S. Tomé e Príncipe.
• Instituto Camões.
• NLX - Grupo de Fala e Linguagem Natural do Departamento de Informática, Universidade de Lisboa, Faculdade de Ciências.
• Universidade Aberta.

Referências bibliográficas

Bacelar do Nascimento, M.F. e 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. e 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. e Amaro, R. (2004). "Morphlogical Tagging of a Spoken Portuguese Corpus Using Available Resources", in Branco, A., Mendes, A. e 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 Kolhapur 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.