Vocabulário Ortográfico Comum (VOC): constituição de uma base lexical para a língua portuguesa

(Common Orthographic Vocabulary (VOC): development of a lexical database for the Portuguese language)

Gladis Maria de Barcellos Almeida[1], José Pedro Ferreira[2], Margarita Correia[3], Gilvan Müller de Oliveira[4]

[1] Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
[2] Instituto de Linguística Teórica e Computacional (ILTEC)
[3] Instituto de Linguística Teórica e Computacional (ILTEC) e Universidade de Lisboa (UL)
[4] Instituto Internacional da Língua Portuguesa (IILP)

gladis@ufscar.br, jpf@iltec.pt, mcf@iltec.pt, gimioliz@gmail.com

Abstract: The International Institute for the Portuguese Language, under the auspices of the Community of Portuguese Speaking Countries, is coordinating the development of the Common Orthographic Vocabulary (VOC). VOC will be a large on-line lexical database which will take into account the varieties of the eight Portuguese speaking countries (Angola, Brazil, Cape Verde, Guinea-Bissau, Mozambique, Portugal, São Tome and Principe, and East Timor). The project comprises two inter-related phases: 1) the merging of existing Portuguese and Brazilian vocabularies as an evidence of the lexicographic tradition; 2) the compilation of corpora for the development of national vocabularies for the remaining countries. This paper describes in detail the methodology of the tasked being pursued in the context of that project.

Key-words: lexical database; corpus; lexical resource; lexicographic tradition; orthographic vocabulary

Resumo: O Instituto Internacional da Língua Portuguesa (IILP), sob os auspícios da Comunidade dos Países de Língua Portuguesa, lidera a organização do Vocabulário Ortográfico Comum da Língua Portuguesa (VOC). O VOC constituir-se-á numa grande base lexical *on-line*, que contemplará as variedades dos oito países lusófonos (Angola, Brasil, Cabo Verde, Guiné-Bissau, Moçambique, Portugal, São Tomé e Príncipe e Timor-Leste). Esse projeto está dividido em duas fases: 1) a junção de vocabulários já existentes em Portugal e Brasil, evidenciando a tradição lexicográfica portuguesa; 2) a elaboração de *corpora* para a constituição dos vocabulários nacionais dos demais países. Assim, nesta oportunidade, pretende-se detalhar os aspectos metodológicos que subjazem às tarefas envolvidas no referido projeto.

Palavras-chave: base lexical; *corpus*; recurso lexical; tradição lexicográfica; vocabulário ortográfico

Introdução

A constituição do Vocabulário Ortográfico Comum da Língua Portuguesa (VOC) é um desiderato da Comunidade dos Países de Língua Portuguesa[1] (CPLP) e uma determinação do Acordo Ortográfico da Língua Portuguesa de 1990[2] (AO90). O projeto

1 Mais informações em <http://www.cplp.org/>.

2 O Acordo Ortográfico da Língua Portuguesa foi assinado em Lisboa, em 16 de dezembro de 1990, por todos os países membros à época. A constituição de um Vocabulário Ortográfico Comum foi um requisito do Acordo: "Os Estados signatários tomarão, através das instituições e órgãos competentes, as providências necessárias com vista à elaboração, até 1 de Janeiro de 1993, de um vocabulário ortográfico comum da língua portuguesa, tão completo quanto desejável e tão normalizador quanto possível, no que se refere às terminologias científicas e técnicas." (Artigo 2.º do texto do AO90).

encontra-se atualmente em curso, sob a coordenação do Instituto Internacional de Língua Portuguesa[3] (IILP). Tentativas anteriores de elaborar um vocabulário ortográfico comum esbarraram nas diferentes interpretações dadas ao texto das reformas ortográficas que pretendiam aplicar, as quais se refletiram no desenvolvimento de vocabulários ortográficos nacionais com divergências (Brasil e Portugal). Nesse contexto, o desenvolvimento do VOC nas condições descritas neste trabalho constitui não só a base da aplicação do AO90 em todos os países da CPLP, como um avanço sem precedentes na lexicografia de língua portuguesa.

A mais recente reforma ortográfica da língua portuguesa, em curso desde 2009,[4] que visa à unificação da ortografia nos vários países da CPLP, apenas unificou as regras de escrita, que preveem a coexistência de variantes nacionais em determinados contextos, e não exige a fusão das tradições lexicográfica e ortográfica. Isto é, continuarão a existir casos de divergência que dependem não da aplicação de regras, mas da tradição, como *úmido* no Brasil e *húmido* nos restantes países. Outro dos problemas que se colocam deve-se ao fato de os diferentes países se encontrarem em momentos diferentes da solidificação e normalização das suas variedades nacionais.

Por essas razões, a solução adotada no VOC é uma abordagem mista, na qual a tradição lexicográfica e o uso real em contextos escritos da variedade padrão são levados em conta para obter uma fatia representativa do léxico da língua. Por um lado, a tradição desempenha um papel fundamental na determinação das formas existentes e na das que não são alteradas pela reforma ortográfica; por outro lado, para que o VOC seja representativo do português como um todo, língua pluricêntrica de oito países, deve registrar, pelo menos, as formas mais frequentes em uso em todos os países, especialmente aqueles com menor ou inexistente tradição lexicográfica.

Assim, o presente artigo pretende detalhar todos os procedimentos envolvidos na elaboração do VOC, desde a junção dos vocabulários já existentes em Portugal e Brasil, pondo em evidência a tradição lexicográfica portuguesa e os aspectos computacionais atinentes à tarefa, como também a constituição dos *corpora* nos demais países membros, retratando a sua importância em projetos envolvendo o léxico.

A junção dos Vocabulários de Portugal e do Brasil

O português dispõe já de uma longa tradição lexicográfica, construída, sobretudo, nos últimos dois séculos em Portugal e no Brasil, que deu origem a um acervo extenso, ainda que não comparável ao de outras línguas de relevância mundial. Por isso, o VOC será constituído por entradas com duas origens, que representarão duas partes distintas da obra: uma, correspondente à *memória lexicográfica do português*, composta pelas entradas dos vocabulários ortográficos oficiais já existentes; outra, correspondente ao léxico obtido através da mineração de *corpora*, representativa do léxico efetivamente em uso em todos os países.

---

3 Instituição vinculada à CPLP, que tem como objetivos "a promoção, a defesa, o enriquecimento e a difusão da língua portuguesa como veículo de cultura, educação, informação e acesso ao conhecimento científico, tecnológico e de utilização oficial em fóruns internacionais" (mais informações em <http://www.iilp.org.cv/>).

4 No Brasil, o AO90 foi promulgado pelo decreto n. 6.583, de 29/9/2008.

A memória lexicográfica do português será constituída por um subconjunto equilibrado das entradas do *Vocabulário Ortográfico do Português*[5] (VOP), sob a responsabilidade de Margarita Correia, produzido pelo Instituto de Linguística Teórica e Computacional[6] (ILTEC) e oficial em Portugal, e do *Vocabulário Ortográfico da Língua Portuguesa*[7] (VOLP), 5ª ed., coordenada por Evanildo Bechara, da Academia Brasileira de Letras. Ambas as obras são baseadas nas nomenclaturas de dicionários existentes,[8] ainda que os *corpora* venham a desempenhar um papel importante na integração de ambos no VOC, como se verá adiante.

Em qualquer obra dessa natureza, que se proponha a integrar informação provinda de várias fontes lexicográficas e com características distintas, uma das principais tarefas passa pela homogeneização das entradas e pelo controle da representatividade de cada uma delas, equilibrando o resultado final. Um primeiro problema que a execução do VOC levanta deve-se ao fato de que as obras que integram a memória lexicográfica do português têm características diferentes quanto aos critérios lexicográficos, de inclusão e de identidade lexical, além de diferentes dimensões e propriedades (o VOLP tem uma nomenclatura muito mais extensa que o VOP, mas inclui muito menos informação para cada entrada). Esse fato impossibilita, portanto, a integração direta das duas obras lexicográficas.

Em face desses problemas, o VOC foi implementado numa plataforma centralizada de edição lexicográfica, que permite a homogeneização das obras, recorrendo a *corpora* para selecionar subconjuntos de dados. Foi escolhida pelo IILP a plataforma usada pelo ILTEC na construção do VOC, o OSLIN[9] (JANSSEN, 2005). Nesse sentido, o VOLP tem de ser inteiramente adaptado a essa plataforma.

O OSLIN (sigla de *Open Source Lexical Information Network*) é uma plataforma digital com base em serviços alojados na web que permite a rápida criação de recursos lexicais, a sua gestão e a permanente inserção de dados. A plataforma contém ferramentas dedicadas a diferentes funções,[10] pensadas para serem usadas por lexicógrafos, que facilitam a gestão e a criação, de forma integrada, a partir de um par *lema – classe gramatical*, de entradas lexicais com outra informação formal associada: paradigma flexional, todas as formas flexionadas, divisão silábica (para fins de translineação), acentuação, remissões para entradas relacionadas (entre elas as variantes diatópicas decorrentes da aplicação do AO90, muito relevantes no caso desta obra), relações funcionais com outras entradas (por ex., todos os nomes deverbais eventivos, nomes de qualidade e advérbios deadjetivais são associados explicitamente) e, dentro em breve, informação sobre a morfologia e a fonética (ASHBY; FERREIRA, 2010).

A integração do VOLP nessa plataforma está sendo feita por uma equipe coordenada por Gladis M. de Barcellos Almeida, na Universidade Federal de São Carlos (UFSCar).

---

5 Disponível em <www.portaldalinguaportuguesa.org>.

6 Conferir em <www.iltec.pt/>.

7 Conferir em <www.academia.org.br>.

8 Para uma descrição do VOP, consulte-se Correia; Ferreira [no prelo]. As fontes do VOLP não são referidas explicitamente na obra, embora, em comunicações públicas dos seus autores, seja ponto assente que teve como base obras lexicográficas do português.

9 Mais informações em <www.oslin.org>.

10 Para uma descrição mais detalhada do sistema multivalente de edição e manutenção dos recursos lexicais do OSLIN, consulte-se Ferreira et al. (2008).

A nomenclatura daquela obra foi primeiro importada para o sistema OSLIN, seguindo operações de equivalência, desdobramento e sobreposição de classes gramaticais, tornando-as (e às entradas que as contêm) compatíveis com as do VOC. Cada entrada do VOLP, agora já adaptada aos critérios de identidade lexical e ao *tagset* do OSLIN, foi depois marcada como estando ou não presente também no VOP, sendo a sua inclusão no sistema validada automaticamente, nos casos em que isso se verificava. O restante do material lexical do VOLP está sendo sistematicamente cruzado e integrado no VOC com base em interseções entre a obra e *corpora* existentes, de modo a finalizar essa fase e a validar os dados.

Em primeiro lugar, o VOLP foi cruzado com um léxico[11] computacional construído a partir do *corpus* do Núcleo Interinstitucional de Linguística Computacional[12] (NILC) (PINHEIRO; ALUÍSIO, 2003). Esse léxico foi primeiramente, tal como o VOLP, adaptado à estrutura lexical e categorial do VOC, depois de conformado com o Acordo Ortográfico, tarefa executada primeiramente utilizando-se o conversor *Lince*[13] (FERREIRA et al., 2012) e, depois, valendo-se de verificações manuais por padrão através de expressões regulares correspondentes aos contextos em que no português do Brasil existem potenciais mudanças.

O VOLP também está sendo cruzado com o *Corpus Brasileiro*[14] versão 1 (CEPRIL/PUCSP/FAPESP) (BERBER SARDINHA et al., 2009), que, além das operações indicadas acima, é adicionalmente necessário proceder à lematização dos *tokens* constantes do *corpus*, à sua etiquetação morfossintática e à sua contagem e ordenação por frequência (o léxico do NILC já tinha sido alvo dessas operações, razão pela qual não foram, nesse caso, necessárias).

Depois dessas operações preparatórias, que permitiram a criação de bases de dados lexicais MySQL independentes para cada fonte, foi executado para cada uma delas o mesmo pré-processamento computacional para os pares *lema – classe gramatical* que foi feito para o VOLP, usando ferramentas disponíveis no sistema de administração do OSLIN, de modo a tornar os dados compatíveis com os das outras bases de dados lexicais. A base de dados contendo o VOLP já de acordo com o formato OSLIN é então fornecida, através de cruzamento, com informação de frequência obtida a partir dos *corpora*, de modo a obter um subconjunto de mais de 200 000 entradas do VOLP que estejam atestadas com suficiente frequência nos *corpora* de referência usados. As entradas a serem inseridas no VOC são, pois, sempre atestadas na fonte lexicográfica de referência, o VOLP, e em *corpora*.

O uso de *corpora* para constituição de novos recursos

Para as restantes variedades do português, no entanto, não é possível adotar o procedimento acima descrito, dado que não existem ainda recursos lexicográficos de referência que as representem.

---

11 Esse léxico é o que dá suporte ao corretor ortográfico do *Microsoft Word*, versão brasileira.

12 Conferir em <www.nilc.icmc.usp.br>.

13 O *Lince* é gratuito, de distribuição livre e pode ser obtido no *site* do Portal da Língua Portuguesa <www.portaldalinguaportuguesa.org >.

14 Conferir em < http://corpusbrasileiro.pucsp.br>.

Seguindo as decisões no seio do IILP, dos representantes de todos os países da CPLP, optou-se pela elaboração de *corpora*. Cada Estado Membro da CPLP assegurou a existência de uma equipe nacional com, no mínimo, três indivíduos que executarão, em coordenação com uma equipe central do projeto, as tarefas necessárias ao prosseguimento do objetivo final.

Um primeiro problema a enfrentar para que isso fosse possível relacionava-se com a inexistência sequer de *corpora* de referência representativos dessas variedades. O pouco trabalho feito nesse sentido refere-se a variedades sociais que não se coadunam com o caráter dos recursos que devem ser criados, que se pretendem representativos de variedades diafasicamente relacionadas com a escrita em contexto formal e com a norma de cada país.

Uma primeira e essencial tarefa do projeto era, por isso, a constituição de *corpora* de referência para cada país da CPLP que deles não dispõe. Como se justifica no decorrer deste artigo, isso deverá ser feito atendendo à representatividade, ao balanceamento, à diversidade e ao tamanho, levando ainda em consideração a exequibilidade e os custos, essenciais para permitir a rápida execução do projeto. Assim, foi elaborado e aprovado pelos representantes técnicos dos vários países um conjunto de metas quanto a tipos textuais, sua proveniência e o peso no resultado final.

Cada *corpus* nacional terá no mínimo 30 milhões de palavras, distribuídas por sub-*-corpora* com texto literário (20%), jornalístico (25%), legislativo e de sessões parlamentares (25%), técnico (saúde, educação, ambiente, pescas e agricultura: 25%) e de proveniência variada (5%). Essa distribuição assegura que todos os países possam ter as mesmas fontes, todas passíveis de serem obtidas em formato digital, de modo a reduzir o tempo de execução e o custo com recursos humanos que seria necessário para processar texto noutros suportes. Além disso, garantiu-se que o material necessário existe em todos os países, o que determinou os domínios de especialidade do sub-*corpus* técnico indicados acima.

Os *corpora* são obtidos por cada equipe nacional em formato digital, renomeados de acordo com o tipo e a origem da fonte e colocados numa estrutura de pastas hierarquizadas num servidor comum. Cada arquivo é depois processado recursivamente pela equipe central, que o converte para texto simples e sugere elementos não textuais para limpeza, e pela equipe nacional, que procede à validação desse primeiro processamento. Posteriormente, cada texto é dividido em extratos, que são anotados quanto à proveniência e tipo, *tokenizado*, lematizado, etiquetado e finalmente o seu material lexical é convertido num léxico de frequência contendo lema, classe de palavra e índice de frequência.[15]

Cada um desses léxicos de frequência é depois inserido numa base de dados própria no sistema OSLIN, tornando-se uma lista de candidatos à inserção no VOC, e, depois de definida uma linha de corte com base na frequência e distribuição por sub-*corpus*, as entradas são inseridas finalmente no VOC por cada equipe nacional usando a plataforma de gestão do sistema. Para isso, foi integrado na plataforma de gestão do OSLIN um módulo de inserção em massa de entradas lexicais. Esse módulo permite gerar automaticamente todas as propriedades formais de cada palavra a partir do lema e da classe gramatical de cada candidato. Os lexicógrafos da equipe nacional de cada país, tal como ocorreram com as entradas obtidas a partir do VOP e do VOLP, verificam manualmente a legitimidade de toda e

15 Para uma descrição detalhada do processo seguido, consultar Almeida e Ferreira (2012).

qualquer entrada e a correção das informações geradas automaticamente pelo sistema, definindo apenas o paradigma flexional de cada entrada, que não pode ser corretamente aduzido pelo sistema.[16]

Cada palavra constante do VOC é marcada explicitamente quanto à sua proveniência e quanto às fontes e países em que se encontra atestada e, nos casos em que esteja disponível, do seu índice de frequência e das variantes já atestadas no Vocabulário.

Está em desenvolvimento um módulo adicional do OSLIN que permite a identificação semiautomática de variantes, que facilitará a difícil tarefa de registrar todas as remissões correspondentes à variação nacional existente entre países, quer a resultante da aplicação das regras do Acordo Ortográfico, nos casos em que permite variação, quer a que não decorre da sua aplicação, dado que o Acordo Ortográfico não unifica a grafia do português, mas apenas a enunciação das suas regras.

O papel dos *corpora* para a avaliação e o estabelecimento de um padrão

A utilização de *corpora* ou grandes quantidades de textos armazenados e selecionados de acordo com determinados critérios para servir de fonte de coleta e/ou análise de itens lexicais não é prerrogativa do mundo moderno. Há muito que se constituem *corpora* para a validação ou definição de uma norma lexical que, na tradição lexicográfica portuguesa e inglesa – para exemplificar com duas grandes línguas pluricêntricas –, viria a ser organizada e publicada em forma de grandes dicionários de língua.

Na língua portuguesa, pode-se citar o *Vocabulário Portuguez e Latino*, elaborado pelo Padre Rafael Bluteau e publicado em oito volumes, entre 1712-1728. O Vocabulário foi o primeiro dicionário para o qual foi fixado um *corpus* (MURAKAWA, 2001). Esse *corpus*, contendo 406 obras, aproximadamente, com autores dos séculos XV a XVII, foi utilizado como exemplário de uso linguístico para as palavras que constavam da nomenclatura do dicionário (MURAKAWA, 2001; 2006). Outro exemplo já no século XIX é o *Diccionario da Lingua Portugueza*, 2ª edição, de António de Morais Silva, publicado em 1813, o qual também se valeu de um *corpus* (MURAKAWA, 2006) como fonte para a recolha de exemplos.

Na língua inglesa, pode-se referir *A Dictionary of the English Language*, publicado em 1755, para o qual Samuel Johnson reuniu um *corpus* de textos para poder observar as palavras em seu uso autêntico. Esse *corpus* lhe serviu também como fonte de exemplos para os verbetes. Outra importante obra é *The Oxford English Dictionary* (OED), cujo projeto lexicográfico, capitaneado por James Murray, também se valeu de uma grande quantidade de textos, obtidos a partir de uma carta de apelo enviada pelo próprio Murray em 1879, pedindo colaborações. Como resposta a essa carta, mais de 800 voluntários leitores passaram a enviar ao editor tiras de papel com citações que deveriam conter as palavras a eles atribuídas (BIBER; CONRAD; REPPEN, 1998). A reunião desses milhares de tiras de papel constituiu o primeiro *corpus* do OED.

Além de os *corpora* terem um importante papel em grandes projetos lexicográficos orientados por uma metodologia empírica, considera-se que a palavra seja a principal unidade de análise dos estudos realizados em Linguística de *Corpus*, dada a facilidade de sua

16 Para uma descrição detalhada deste módulo, consultar Janssen (2011).

identificação pelas ferramentas computacionais (VIANA, 2011). Em projetos envolvendo o léxico, a palavra é praticamente a porta de entrada para a análise de *corpus*, ou, como afirma Calzolari (1996, p. 3), "all Language Engineering applications require knowledge about words". Não é sem razão, pois, que projetos e/ou atividades de pesquisa situados no âmbito do léxico tenham no *corpus* seu maior aliado.

O que mudou daqueles projetos lexicográficos fundadores para cá é a concepção de *corpus*, entendido hoje como "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (SINCLAIR, 2005, p. 10). Esse formato eletrônico, promovido pelo advento do computador, interferiu diretamente não só na concepção que se tem de *corpus* como também na sua forma de armazenamento e exploração, já que os recursos oferecidos pela máquina permitiram que grandes quantidades de textos pudessem ser processadas em questão de segundos, fazendo com que muitas hipóteses sobre determinados fenômenos linguísticos pudessem ser testadas rápida e eficientemente.

A moderna noção de *corpus* também carrega consigo requisitos que devem ser fortemente considerados num projeto de *corpus*. São eles: representatividade, balanceamento, diversidade e tamanho (McENERY; WILSON, 1996; KENNEDY, 1998; BIBER et al., 1998; RENOUF, 1998; SINCLAIR, 2005).

Dentre todos esses requisitos, a representatividade é crucial, haja vista que um *corpus* representativo tende a ser bem balanceado, ter boa diversidade e tamanho adequado aos objetivos da pesquisa. Assim, para a construção do VOC, especial atenção foi dada à representatividade, já que o vocabulário ortográfico deverá servir como importante instrumento de normalização lexical da língua portuguesa.

Quanto ao caso específico das variedades menos representadas do português, há que se levar em conta a discrepância entre as condições linguísticas da língua no conjunto dos oito países membros da CPLP: Angola, Brasil, Cabo Verde, Guiné-Bissau, Moçambique, Portugal, São Tomé e Príncipe e Timor-Leste.

Com exceção de Portugal e Brasil, que já têm *corpora*, grandes dicionários tradicionais (instrumentos linguísticos que permitem a construção e são muitas vezes a base de vocabulários ortográficos) e são países onde o português é majoritariamente a língua mais usada, os demais países (Angola, Cabo Verde, Guiné-Bissau, Moçambique, São Tomé e Príncipe e Timor-Leste) vivem uma situação linguística bastante diferente, uma vez que grande parte dos seus cidadãos não tem o português como primeira língua.

Observe-se, no Quadro 1, a quantidade de línguas que convive com o português nos Países Africanos de Língua Oficial Portuguesa (PALOP) e em Timor-Leste, numa lista necessariamente incompleta dada a inexistência de consenso em relação a cada país e, em alguns casos, à falta de dados e de estudos recentes. Apresenta-se, também, no mesmo Quadro, a população aproximada de cada um e o ano de sua independência, para se ter uma noção um pouco mais ampla do cenário linguístico, demográfico e político no qual se insere este projeto do VOC.

Quadro 1: Línguas faladas nos PALOPs e em Timor-Leste, além do português[17]

Estados-membros	Línguas faladas (além do português)	População	Ano da independência
Angola	*Kikongo*, *kimbundo*, *tchokwe*, umbundo, mbunda, *kwanyama*, nhaneca, fiote, nganguela, entre outras	18.498.000	1975
Cabo Verde	Crioulo de Cabo-Verde (ou caboverdiano)	499.796	1975
Guiné-Bissau	Crioulo da Guiné-Bissau (ou *kriol*), balanta, mandjaco, fula, entre outras	1.449.230	1974
Moçambique	*Cicopi*, *cinyanja*, *cinyungwe*, *cisenga*, *cishona*, *ciyao*, *echuwabo*, *ekoti*, *elomwe*, *gitonga*, maconde (ou *shimakonde*), *kimwani*, macua (ou *emakhuwa*), memane, suaíli (ou *kiswahili*), suazi (ou *swazi*), xichangana, *xironga*, *xitswa* e zulu	20.366.795	1975
São Tomé e Príncipe	Forro, principense (ou *lunguyé*) e angolar.	187.356	1975
Timor-Leste	Tétum, entre muitas outras línguas austronésias e algumas papuanas	1.066.582	1975[18] e 2002[19]

Tendo em vista, pois, esse cenário, antes de se pensar em constituir vocabulários nacionais de cada um desses países, é preciso primeiro constituir os *corpora* que representem, de fato, a norma lexical de cada um deles; *corpora* que incluam tanto palavras comuns a todos ou a alguns países como aquelas que são específicas de apenas um deles.

Além desse aspecto, foi preciso ter em conta que os *corpora* compilados fossem comparáveis em tamanho, em distribuição por gêneros e em relação ao peso de cada um dos gêneros no conjunto final (ALMEIDA; FERREIRA, 2012), de modo que o VOC, depois de pronto, evidenciasse uma gestão política de língua descentralizada, despolarizada (saindo do eixo Portugal e Brasil) e com participação ativa de todos os países-membros.

Critérios iniciais que orientaram o projeto dos *corpora* foram: i) textos escritos e em contexto formal; ii) prioridade para textos já em formato digital. A partir daí, iniciou-se um trabalho coletivo, contando com a cooperação das equipes nacionais de cada país, de forma a estabelecer as características dos *corpora*. A partir dessas primeiras decisões, foram-se desenhando os *corpora*, de modo a atender o principal requisito, qual seja, a representatividade.

Ser **representativo** significa que um *corpus* deve ser elaborado de forma a espelhar determinadas características linguísticas da comunidade cuja língua está sob análise (SINCLAIR, 2005). Daí a importância de se fazerem escolhas adequadas para que o *corpus* possa realmente refletir comportamentos linguísticos. Questões que devem ser feitas durante a seleção dos textos são: quais documentos? Quais tipos de textos? Quais gêneros textuais? Enfim, o que de fato representa os usos linguísticos de uma comunidade? No

17 As informações foram obtidas no portal da CPLP (http://www.cplp.org/); nos portais dos respectivos governos (Angola - http://www.governo.gov.ao/; Cabo Verde - http://www.governo.cv/; Guiné-Bissau - http://www.anpguinebissau.org/; Moçambique - http://www.portaldogoverno.gov.mz; São Tomé e Príncipe - http://www.presidencia.st/; Timor-Leste http://timor-leste.gov.tl), na Wikipédia, especificamente para conferir os dados populacionais e em Batoréo e Casadinho (2009).

18 Independência de Portugal.

19 Fim da ocupação indonésia.

caso dos *corpora* para o VOC, a pergunta teve de ser ainda mais específica: quando se escreve em português num contexto formal nesses países, quais palavras são empregadas?

Diretamente associado à representatividade, está o conceito de **balanceamento**. Embora seja um conceito vago, de acordo com Sinclair (2005), é preciso ter em mente que o *corpus* deve ter um equilíbrio entre gêneros discursivos (informativo, científico, etc.), tipos de textos incluídos (artigo, editorial, entrevista, dissertação, etc.), temas (pesca, agricultura, saúde, educação, etc.), ou até mesmo títulos, ou autores. O ideal seria que se conseguisse levar em conta todas essas categorias, mas sempre atendendo às demandas da pesquisa que se pretende realizar.

Se o *corpus* for cuidadosamente balanceado, ele terá uma boa **diversidade**. A propósito disso, Biber et al. (1998, p. 248) assinalam que "there are important differences in the use of lexical, grammatical and discourse features across different varieties of language". O que lhe dá argumentos para afirmar que o conceito de "língua geral" é hipotético, dado que cada gênero discursivo tem seus próprios padrões de uso. Nesse sentido, um *corpus* representativo deve conter uma diversidade de gêneros, tipos de textos e assuntos, pois a frequência de muitas palavras pode variar de acordo com o assunto (BIBER et al., 1998).

Um *corpus* pensado para ser representativo, balanceado e diversificado tem, em geral, um **tamanho** adequado ao tipo de pesquisa que se vai realizar e à metodologia a ser adotada na pesquisa (SINCLAIR, 2005). Quando se fala em tamanho de *corpus*, não se trata somente do número total de ocorrências (*tokens*) e de palavras diferentes (*types*), mas com quantas categorias (gêneros discursivos, tipos de textos, assuntos, títulos, autores, etc.) um *corpus* deve contar, quantas amostras de cada categoria e quantas palavras existem dentro de cada amostra (KENNEDY, 1998).

Assim, os *corpora* que estão sendo compilados nos PALOPs e em Timor-Leste atendem a esses requisitos de representatividade, balanceamento, diversidade e tamanho, acima especificados.

Todavia, para a gestão descentralizada de uma língua pluricêntrica, tão importante quanto esses requisitos é o envolvimento de todas as equipes nacionais no projeto; afinal, quem mais sabe responder às questões que devem ser feitas durante a seleção de textos para compor um *corpus* são os falantes de cada um desses países.

O uso de *corpora* no VOC

Tendo em conta o que foi dito, os *corpora* assumem necessariamente um papel central na criação de novos recursos lexicográficos para a língua portuguesa quando tomada como um todo, como é o caso do VOC.

Por um lado, pela sua natureza, herança e princípios metodológicos, os dicionários existentes para o português são mais um perpetuar da tradição lexicográfica, ainda que em alguns casos parcialmente atualizada, do que um verdadeiro repositório atualizado das palavras de fato em uso nos países de língua portuguesa, o que afeta a sua representatividade. Por isso, uma primeira função dos *corpora* no VOC é a de validação dos dados constantes da tradição lexicográfica portuguesa.

Além disso, dado que a reforma ortográfica não torna homogênea a escrita de muitas palavras com formas ortográficas já antes divergentes (*úmido* no Brasil, mas *húmido* nos

restantes países, como já mencionamos no início deste artigo), qualquer repositório lexical multinacional para o português deve identificar, marcar e relacionar por meio de remissões as diversas variantes, o que só é possível executar num curto espaço de tempo recorrendo à análise semiautomática das divergências entre *corpora* representativos de cada variedade.

Por último, atendendo à falta de cobertura lexicográfica atual para muitas das variedades nacionais do português e à inexistência de *corpora* representativos dessas variedades, a criação e o uso de tais recursos de base são essenciais para captar a realidade da língua escrita em contexto formal em cada país e proceder à sua integração num recurso que a represente, fixando a ortografia do seu léxico. Nesse sentido, os *corpora* servem, pois, de fonte primária para as entradas da obra em criação.

Considerações finais

Apesar de todo o trabalho desenvolvido, sobretudo no último século e meio, e da grande evolução trazida pelas publicações mais recentes de dicionários, a lexicografia de referência do português tem ainda claras lacunas face às outras línguas de relevância mundial. São particularmente prementes três problemas: a falta de recursos lexicais normalizadores disponíveis que permitam o processamento computacional da língua portuguesa; a falta de recursos feitos com base ou tendo em conta a informação obtida a partir de *corpora*; a falta de recursos representativos da diversidade do português, que possam agir nacionalmente como normalizadores em países que não dispõem neste momento de recursos próprios.

Além dessas questões puramente lexicográficas, o português carece, além disso, de um vocabulário ortográfico comum, prerrogativa do Acordo Ortográfico firmado por todos os países e representante da vontade política manifestada de manter a unidade da língua portuguesa.

O VOC pretende responder a esses problemas, criando um novo recurso multivalente que reaproveita os recursos já disponíveis e cria, quando necessário, novos recursos. O projeto, ainda em fase de desenvolvimento, dará os seus primeiros resultados em 2014. Os resultados serão, a partir dessa data, disponibilizados gratuitamente a partir de uma interface *on-line* de acesso livre.

REFERÊNCIAS

ALMEIDA, G. M. B.; FERREIRA, J. P. *Manual para a elaboração de corpora.* Com vista à organização dos Vocabulários Ortográficos Nacionais dos países integrantes da CPLP. Lisboa (Portugal): ILTEC; São Carlos (SP, Brasil): NILC, 2012.

ASHBY, S.; FERREIRA, J. P. The Role of Morphology in Generating High-Quality Pronunciation Lexica for Regional Variants of Portuguese. In: BRANCO, A.; KLAUTAU, A.; VIEIRA, R.; LIMA, V. L. S. de; PARDO, T. A. S. (Ed.). *Computational Processing of the Portuguese Language* – Lecture Notes in Artificial Intelligence, v. 6001. Berlin Heidelberg: Springer-Verlag, 2010. p. 162-165.

BATORÉO, H.; CASADINHO, M. O português, uma língua pluricêntrica: o caso de Timor-Leste. *Revista Portuguesa de Humanidades*, Estudos Linguísticos, Braga: Universidade Católica Portuguesa de Braga, 13-1, p. 63-79, 2009. Disponível em: <http://www.catedraportugues.uem.mz/lib/docs/bib_timor/Batoreo_Casadinho_2009.pdf>. Acesso em: 12 ago. 2012.

BERBER SARDINHA, T.; MOREIRA FILHO, J. L.; ALAMBERT, E. The Brazilian *Corpus*: A one-billion word online resource. In: MAHLBERG, M.; GONZÁLEZ-DÍAZ, V.; SMITH, C. (Ed.). *Proceedings of the Fifth Corpus Linguistics Conference*, CL2009. Liverpool: University of Liverpool, UK, 20-23 July 2009.

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics* – Investigating Language Structure and Use. Cambridge, U.K.: Cambridge University Press, 1998.

CALZOLARI, N. Lexicon and *Corpus*: a multi-faceted interaction. In: GELLERSTAM, M.; JARBORG, J.; MALMGREN, S.-G.; NOREN, K.; ROGSTROM, L.; PAPMEHL , C. R. (Ed.). *Euralex '96 Proceedings*. Goteborg: Göteborgs Universitet, 1996. p. 3-16.

CORREIA, M.; FERREIRA, J. P. Vocabulário Ortográfico do Português Correia, Margarita e Ferreira, José Pedro. In: *Actas del V Congreso Internacional de Lexicografía Hispánica*, Madrid, 25-27 jun. 2012 (no prelo)

FERREIRA, J. P.; BARBOSA, S.; JANSSEN, M. Mordebe Admin: A Lexical Management System. In: *Proceedings of the 13th Euralex International Congress, Barcelona, 2008*. Barcelona: Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra; Documenta Universitaria, 2008. p. 351-357.

FERREIRA, J. P.; LOURINHO, A.; CORREIA, M. Lince, an End User Tool for the Implementation of the Spelling Reform of Portuguese. In: CASELI, H. M.; VILLAVICENCIO, A.; TEIXEIRA, A. J. S.; Perdigão, F. (Ed.). *Computational Processing of the Portuguese Language* - 10th International Conference, PROPOR 2012 Proceedings. Lecture Notes in Computer Science 7243. Berlin, Heidelberg: Springer, 2012. p. 46-55.

JANSSEN, M. Open Source Lexical Information Network. In: BOUILLON, P.; KANZAKI, K. (Ed.). *Proceedings of the Third International Workshop on Generative Approaches to the Lexicon*, May 19-21 2005. Genebra: École de Traduction et d'Interpretation – Université de Genéve, 2005. p. 79-106.

_____. Computer-Aided Inflection for Lexicography Controlled Lexica. In: KOSEM, I.; KOSEM, K. *Electronic Lexicography in the 21st Century New Applications for New Users* – Proceedings of eLex 2011. Liubliana: Trojina, Institute for Applied Slovene Studies, 2011. p. 96-105.

KENNEDY, G. *An Introduction to Corpus Linguistics*. London/NY: Longman, 1998.

McENERY, T.; WILSON, A. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 1996.

MURAKAWA, C. A. A. Tradição lexicográfica em língua portuguesa. In: OLIVEIRA, A. M. P.; ISQUERDO, A. N. (Ed.). *As ciências do léxico*: lexicologia, lexicografia e terminologia. 2. ed. Campo Grande: Ed. UFMS, 2001. p. 153-159.

_____. *António de Morais Silva*: lexicógrafo da língua portuguesa. Araraquara: Laboratório Editorial FCL/UNESP; São Paulo: Cultura Acadêmica Editora, 2006. 228p.

PINHEIRO, G. M.; ALUÍSIO, S. M. *Córpus Nilc*: descrição e análise crítica com vistas ao projeto Lacio-Web. NILC-TR-03-03, 2003. Disponível em: <http://www.nilc.icmc.usp.br/lacioweb/downloads/NILC-TR-03-03.zip>. Acesso em: 8 set. 2012.

RENOUF, A. (Ed.). *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 1998.

SINCLAIR, J. *Corpus* and Text – Basic Principles. In: Wynne, M. (Ed.). *Developing Linguistic Corpora*: a Guide to Good Practice. Oxford: Oxbow: p. 1-16, 2005. Disponível em: <http://ota.ahds.ac.uk/documents/creating/dlc/chapter1.htm>. Acesso em: 12 ago. 2012.

VIANA, V. Linguística de *corpus*: conceitos, técnicas & análises. In: VIANA, V.; TAGNIN, S. E. O. (Ed.) *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2011. p. 22-92.