Psicologia: Reflexão e Crítica, in press

Procura-PALavras (P-PAL): Uma nova medida de frequência lexical do Português Europeu contemporâneo

Ana Paula Soares[1], Álvaro Iriarte[2], José João de Almeida[3], Alberto Simões[3,2], Ana Costa[1], Patrícia França[1], João Machado[1], & Montserrat Comesaña[1]

[1] Escola de Psicologia, Universidade do Minho, Portugal

[2] Instituto de Letras e Ciências Humanas, Universidade do Minho, Portugal

[3] Departamento de Informática, Universidade do Minho, Portugal

Correspondência relacionada com o artigo deverá ser enviada para:

Ana Paula Soares
Departamento de Psicologia Básica
Escola de Psicologia
Universidade do Minho
Campus de Gualtar
4710-057 Braga
Portugal
Email: asoares@psi.uminho.pt

Resumo

Neste trabalho apresentamos as estratégias e os procedimentos adoptados na constituição de uma nova medida de frequência lexical do Português Europeu contemporâneo, o Procura-PALavras (P-PAL). Baseado num *corpus* de mais de 227 milhões de palavras, o P-PAL é uma aplicação *web* que oferece, por defeito, valores de frequência lexical para todas as suas entradas lexicais (lemas e formas) e que permite a computação de uma grande diversidade de outras medidas objectivas (lexicais e sublexicais) e subjectivas. Descrevemos ainda o processo de definição das suas entradas lexicais e da extracção dos respectivos valores de frequência. O elevado número de índices e de entradas lexicais tornam o P-PAL numa aplicação *web* avançada e indispensável à promoção e internacionalização da investigação em Portugal. O P-PAL encontra-se disponível em http://p-pal.di.uminho.pt/about/tools.

Palavras-chave: FREQUÊNCIA LEXICAL; BASES LEXICAIS; *CORPUS*/*CORPORA*; PORTUGUÊS EUROPEU.

Abstract

In this paper we present the strategies and procedures undertaken in the development of the new frequency lexicon Procura-PALavras (P-PAL). Based on a *corpus* of over 227 million words, P-PAL offers the default word frequency per million words (lemmas and wordforms), and the computation of several other objective (lexical and sublexical) and subjective metrics. We also describe lexical entry integration and word frequency extraction. The number of indices and lexical entries provided makes P-PAL an advanced and indispensable web application for the promotion and internationalization of Portuguese research. P-PAL is available at http://p-pal.di.uminho.pt/tools.

Keywords: WORD FREQUENCY; LEXICAL DATABASES; *CORPUS*/*CORPORA*; EUROPEAN PORTUGUESE.

Introdução

O Procura-PALavras (P-PAL) é um projecto financiado pela Fundação para a Ciência e a Tecnologia (PTDC/PSI-PCO/104679/2008) desenvolvido com o intuito de disponibilizar à comunidade científica uma aplicação *web* com índices psicolinguísticos objectivos e subjectivos de palavras do Português Europeu (PE) contemporâneo. A importância da existência de bases lexicais informatizadas que apoiem de forma efectiva a investigação nas áreas da Psicolinguística, das Neurociências, da Linguística ou da Psicologia Cognitiva em geral é, na actualidade, inquestionável. Com efeito, constituindo a palavra a matéria-prima a partir da qual grande parte da investigação nessas áreas se realiza, e constituindo as palavras, estímulos complexos que reúnem um conjunto de propriedades ou atributos cuja análise, controlo ou manipulação se revelam fundamentais ao desenvolvimento de estudos nesses domínios, a investigação actual já não se compadece mais com a inexistência deste tipo de ferramentas (Soares *et al*., 2010).

Entre essas características, encontram-se tanto propriedades mais objectivas, determinadas pela análise da própria palavra (ex., extensão da palavra em letras ou sílabas, divisão silábica, categoria morfo-sintáctica etc.) ou derivadas da análise da sua relação com as restantes existentes no léxico a nível lexical (ex., similaridade ortográfica ou fonológica com outras palavras) ou sublexical (ex., bigrama, trigrama ou bifone etc.), como propriedades de natureza mais subjectiva que reflectem as experiências pessoais dos indivíduos com o uso da própria língua (ex., imaginabilidade, familiaridade, concreteza etc. - para mais informações sobre os índices e as métricas disponíveis na aplicação ver Soares *et al*., 2010 e/ou consultar http://p-pal.di.uminho.pt/project). A manipulação e/ou o controlo sistemáticos destas propriedades na literatura têm contribuído de forma decisiva não só para a compreensão da arquitectura funcional e o processamento linguístico humano, como para um conhecimento mais aprofundado da própria língua.

Contudo, apesar da relevância de bases como o P-PAL e de elas se encontrarem disponíveis em línguas como o inglês (ex., MRC - Coltheart, 1981; N-Watch – Davis, 2005), o francês (ex., BRULEX - Content, Mousty, & Radeau, 1990; LEXIQUE - New, Pallier, Brysbaert, & Ferrand, 2004), o holandês e o alemão (ex., CELEX - Baayen, Piepenbrock, & Gulikers, 1995), o espanhol (LEXESP – Sebastián-Gallés, Martí, Cuetos, & Carreiras, 2000; BuscaPalabras – Davis & Perea, 2005), o grego (GreekLex - Ktori, van Heuven, & Pitchford, 2008) ou o árabe (ARALEX - Boudelaa & Marslen-Wilson, 2010), elas são escassas no PE. O P-PAL, disponível em http://p-pal.di.uminho.pt/tools, procura colmatar esta necessidade oferecendo à comunidade científica uma aplicação informática multi-plataforma que, com comodidade e rapidez, permite que os investigadores acedam a um vasto conjunto de índices sobre palavras portuguesas, seleccionando, do menu de análises, as métricas que se adeqúem aos propósitos da sua investigação numa dupla possibilidade de análise: (i) obter palavras que obedeçam a determinados requisitos; ou (ii) analisar palavras num conjunto requisitos. De destacar ainda que, em cada uma destas funcionalidades, a aplicação permite realizar pesquisas para lemas ou para formas. Entenda-se por lema o item lexical abstracto escolhido para representar todas as formas flexionadas de uma palavra (ex., "ir" é o lema que representa as formas verbais "vou", "ir", "indo", "fui" e "ido") e por forma a ocorrência natural de uma palavra na língua (ex., "correr", "corre" e "correu" são formas do lema "correr").

Entre a ampla diversidade de medidas a disponibilizar, o P-PAL oferece, por defeito, o valor de frequência lexical para todas as suas entradas (lemas e formas). A frequência lexical é uma medida objectiva que contabiliza o número de ocorrências das palavras numa língua, sendo tradicionalmente obtida através da compilação de grandes quantidades de texto, isto é, da constituição de *corpora*. Um *corpus* pode ser entendido como uma colecção de porções de texto seleccionadas de acordo com um conjunto de critérios para representar, tanto quanto possível, uma determinada língua (Sinclair, 2005), sendo o seu uso para extracção de

frequências uma prática com mais de meio século de existência (ver Sardinha, 2004). Thorndike foi pioneiro nesta abordagem, apresentando em 1944, juntamente com Lorge, uma listagem de frequências de 30,000 palavras extraídas da compilação manual de textos ingleses num total de 4,5 milhões de palavras. Mais tarde, apareceu o primeiro *corpus* electrónico, o *corpus* Brown, a partir do qual se extraíram as normas de frequência mais utilizadas em língua inglesa: as normas de Kučera e Francis (1967). Contudo, apesar da sua popularidade, a validade destas normas tem sido questionada (ex., Balota *et al*., 2004; Brysbaert & Cortese, 2011; Brysbaert & New, 2009). Com efeito, esses valores assentam não só num *corpus* limitado e de pequenas dimensões como desactualizado (≈1 milhão de palavras). Desde então outros *corpora* foram desenvolvidos, sendo de assinalar o *American Heritage Word Frequency* de Carroll, Davis e Richman (1971), cujas frequências derivam de um *corpus* de 5,09 milhões de palavras; o *Educator´s Word Frequency Guide* de Zeno, Ivens Millard e Duvvuri (1995), que contém medidas de frequência extraídas de um *corpus* de 17 milhões de palavras; o *Hyperspace Analogue to Language* (HAL) que se baseou num *corpus* de 131 milhões de palavras (Lund & Burgess, 1996); e, mais recentemente, o *Google Ngram* Viewer (Michel *et al*., 2011) baseado num *corpus* gigantesco de livros publicados desde 1800.

O recurso a este tipo de medidas tem assumido grande importância na investigação. Burgess e Livesay (1998), por exemplo, verificaram que 20% dos artigos publicados entre 1995-1996 em quatro revistas de referência na área da Psicologia Cognitiva (*Journal of Experimental Psychology: Learning, Memory, and Cognition*; *Journal of Memory and Language*; *Memory & Cognition* e *Journal of Psycholinguistic Research*) recorreram a medidas de frequência lexical para o controlo e/ou manipulação de variáveis e esta tendência tem vindo a acentuar-se desde então. Esta situação não é de estranhar se considerarmos que, desde as primeiras observações empíricas realizadas por Cattell (1886), se tem demonstrado de forma sistemática que as palavras mais frequentes são reconhecidas (ex., Forster &

Chambers, 1973; Murray & Forster, 2004), nomeadas (ex., Balota & Chumbley, 1985; Dahan, Magnuson, & Tanenhaus, 2001) e/ou classificadas (ex., Forster & Hector, 2002; Forster & Shen, 1996) de forma mais rápida e precisa do que palavras de baixa frequência. Além disso, estudos recentes (ex., Brysbaert & Cortese, 2011; Brysbaert & New, 2009; Brysbaert *et al*., 2011; Thompson & Desroches, 2009; Zevin & Seidenberg, 2002), têm também comprovado que a frequência de uso das palavras é a variável mais potente na explicação do desempenho linguístico dos sujeitos, capturando mais de 40% da variância dos resultados. Desta forma, como referem Brysbaert *et al.* (2011): "*Because of the importance of word frequency, no study in word recognition or memory research can afford to leave out this variable*" (p. 413).

Contudo, apesar da relevância desta variável, até recentemente não dispúnhamos para o PE de uma medida fiável de frequências. Até ao ano 2000, o único léxico de frequências disponível era o *Português Fundamental* (1984), cujos valores foram extraídos de um pequeno *corpus* oral de pequenas dimensões (700,000 palavras) recolhido nos anos 70. Reconhecendo as limitações deste *corpus*, o Centro de Linguística da Universidade de Lisboa (CLUL) desenvolveu no início dos anos 2000 o CORLEX (Bacelar do Nascimento, Pereira & Saramago, 2000), um léxico com valores de frequência extraídos de um *corpus* de mais de 16 milhões de palavras. Contudo, as indicações mais recentes da literatura alertam para a importância de basear as medidas de frequência em *corpora* de pelo menos 20-30 milhões de palavras (cf. Brysbaert & New, 2009; Brysbaert *et al*., 2011). A sua extracção a partir de *corpora* de menores dimensões pode subestimar a ocorrência das palavras, especialmente as de baixa frequência. Esta situação é tanto mais relevante quanto os trabalhos recentes levados a cabo no âmbito do *English Lexicon Project* (Balota *et al*., 2007), do *French Lexicon Project* (Ferrand *et al*., 2010) e do *Dutch Lexicon Project* (Keuleers *et al*., 2010) revelarem que a quase totalidade do efeito de frequência se situa nos intervalos de frequência abaixo das 10 ocorrências por milhão de palavras. Além disso, como refere Lee (2003), de um ponto de

vista estatístico, a extracção de frequências é mais adequada a partir de grandes amostras porque o erro padrão de medida varia em função da raiz quadrada do tamanho da amostra. Desta forma, a extracção de frequências a partir de *corpora* de maiores dimensões apresenta grandes vantagens, permitindo não só minimizar o erro de medida, como fazer aumentar a probabilidade de palavras de baixa ocorrência se verem representadas no léxico, estabelecendo distinções mais finas e subtis entre elas.

Ora na actualidade o PE conta já com vários léxicos de frequências extraídos de *corpora* de grandes dimensões como os disponibilizados pela Linguateca (Costa, Santos & Cardoso, 2008). No seu projecto *Acesso a corpos/Disponibilização de corpos*, este centro de recursos permite aceder a informações sobre frequências em 19 *corpora* de vários géneros do PE arcaico e contemporâneo e do Português do Brasil. Não obstante a relevância deste projecto e dos recursos que disponibiliza, a pesquisa de frequências apenas pode ser feita em cada um dos *corpora* ou em todos os *corpora* em simultâneo, o que resulta necessariamente numa percentagem de incidências na variante do Português do Brasil e do Português arcaico. Além disso, dado que cada *corpus* apresenta um género específico (ex., jornalístico, técnico, literário) a pesquisa por *corpus* torna o valor de frequência demasiado dependente do seu contexto de extracção. Com efeito, sendo o objectivo deste tipo de medida o de informar acerca da probabilidade de ocorrência das palavras, não no *corpus* de onde são extraídas, mas na língua de onde o *corpus* foi derivado, assume-se como essencial que ele seja o mais diversificado possível na sua composição. A diversidade de género e das modalidades discursivas (oral e escrita) asseguram maior representatividade ao *corpus* e, assim, maior validade às medidas de frequência extraídas a partir dele (Sardinha, 2004; Sinclair, 2005).

Neste contexto, atendendo à relevância da medida de frequência lexical na condução da investigação mais actual, à ausência de uma medida fiável dessa variável para o PE, e à existência recente nessa língua de vários léxicos de frequências extraídos de *corpora* de

grandes dimensões, o presente projecto procura, a partir deles, criar uma nova medida de frequência lexical para o PE contemporâneo. O desenvolvimento de um projecto desta natureza assume-se como de primordial importância dado que não só habilitará os investigadores nacionais com uma ferramenta de valor inestimável à prossecução da investigação nos mais diversos domínios teóricos e aplicados da pesquisa científica – ao apoiar, por exemplo, uma selecção mais eficiente dos estímulos verbais a manipular e/ou controlar – como poderá igualmente concorrer para o desenvolvimento de estudos que permitam um conhecimento mais aprofundado da própria língua – a partir, por exemplo, da análise empírica das características fonológicas, morfo-sintácticas e semânticas do PE contemporâneo. De assinalar ainda o seu potencial contributo para o desenvolvimento de aplicações informáticas mais sofisticadas que, no âmbito do Processamento de Linguagem Natural (PLN) permitam, por exemplo, a construção de dicionários que atendam à vizinhança ortográfica e fonética das palavras, ou de instrumentos de síntese de voz ou de tradução mais eficientes. A aplicação P-PAL assume-se assim tanto como um meio de apoio à investigação em diferentes áreas da pesquisa científica (ex., Psicolinguística, Neurociências, Psicologia Cognitiva em geral), como um objecto de investigação *per se* em domínios tão diversos como o PLN ou a Linguística. Poderá também contribuir para a construção de provas de avaliação (neuro)psicológica que exigem um controlo rigoroso e fiável dos itens a incluir, situando-se portanto aqui mais um contributo a assinalar no desenvolvimento deste tipo de ferramenta.

Em suma, pela inovação que constitui, pela diversidade de índices e métricas que aglutina (para além do novo índice de frequência lexical aqui apresentado, inclui também todo um conjunto de outros índices objectivos - lexicais e sublexicais – e subjectivos de palavras do PE contemporâneo), pela dupla funcionalidade de análises que oferece ao utilizador (avaliar palavras em determinados parâmetros ou obter palavras que obedeçam a tais parâmetros), numa aplicação informática amigável de acesso gratuito, consideramos estar

perante uma ferramenta com um potencial inestimável à promoção e internacionalização da investigação em Portugal.

Método

Não pretendendo o projecto P-PAL criar um novo *corpus* do PE contemporâneo mas antes rentabilizar os *corpora* do PE já existentes procedemos, em primeiro lugar, à identificação dos *corpora* do PE contemporâneo de acesso livre e etiquetados morfossintaticamente para, de seguida, os analisarmos, tratarmos e indexarmos com vista à extracção de uma nova medida de frequência lexical do PE contemporâneo. Neste processo identificámos oito *corpora* (sete disponibilizados pela Linguateca e um pelo CLUL), que passamos a descrever de seguida.

Materiais

O CETEMPúblico é, tanto quanto sabemos, o maior *corpus* do PE disponibilizado gratuitamente pela Linguateca. Constituído por 191,687,833 palavras retiradas de edições do jornal *Público* publicadas entre 1991 e 1998, o CETEMPúblico apresenta informação de frequência para 1,247,135 lemas e 863,933 formas. O Avante! é outro *corpus* jornalístico do PE disponibilizado pela Linguateca. Constituído por 6,501,146 palavras, apresenta valores de frequência para 121,409 formas e 90,081 lemas extraídos de textos do jornal *Avante!* do Partido Comunista Português, de Abril de 1997 até Fevereiro de 2002. O *corpus* DiaCLAV é também um *corpus* do género jornalístico. Elaborado a partir de 12,801 artigos de edições *online* dos jornais regionais o *Diário de Coimbra*, *Diário de Leiria*, *Diário de Aveiro* e *Viseu Diário* datados de Junho de 1999 a Setembro de 2000, é constituído por 6,651,523 ocorrências que originaram 110,063 formas e 86,046 lemas. O Natura/Minho é outro dos *corpora* jornalísticos da Linguateca, desenvolvido pelo grupo de investigação em Processamento de Linguagem Natural da Universidade do Minho. É constituído por textos

retirados de edições do jornal regional *Diário do Minho* de 1999 e integra 1,749,068 ocorrências, 58,956 formas únicas e 57,533 lemas.

Dos *corpora* da Linguateca fazem ainda parte do P-PAL o *corpus* técnico-científico ECI-EE, que contém 27,138 palavras das quais foram extraídas 4,254 formas e 2,719 lemas, e o *corpus* oral Museu da Pessoa, criado a partir de transcrições de entrevistas elaboradas pelo Núcleo Português do Museu da Pessoa. De referir que, muito embora na versão disponibilizada pela Linguateca o *corpus* contenha entrevistas realizadas a falantes do português do Brasil, para o P-PAL considerou-se apenas o registo da variante europeia que contém 362,517 palavras, 21,542 formas e 11,976 lemas. Por fim, foi ainda integrada a parte contemporânea do *corpus* literário Vercial. Com efeito, embora na versão disponibilizada pela Linguateca este *corpus* contenha obras de autores portugueses publicadas entre 1500 e 1933, para o P-PAL foram apenas contabilizadas as obras dos séculos XIX e XX. A parte contemporânea é composta por 4,581,089 palavras, 375,323 formas únicas e 57,533 lemas.

Do CLUL foi integrado o CORLEX (Bacelar do Nascimento *et al.*, 2000). Trata-se de um *corpus* constituído por 16,210,438 ocorrências das quais se extraiu informação de frequência para 26,980 lemas e 140,976 formas. O léxico deriva de um sub*corpus* escrito (15,354,243 palavras) de texto jornalístico, literário, técnico, científico e didáctico e "miscelânea" (que inclui ocorrências oriundas de jornais e revistas especializados e outros documentos), datado entre a segunda metade do século XIX e 1998. As restantes ocorrências (856,195 palavras) derivam de um sub*corpus* oral, constituído pela transcrição do registo magnético de conversas informais e de produções mais formais (conferências, entrevistas de rádio e de televisão, etc.) de 1970 a 1990.

Comparativamente aos restantes *corpora* disponibilizados pela Linguateca que apresentam um único género linguístico, o CORLEX é um *corpus* heterogéneo. Esta situação levantou algumas questões relativamente à sua integração no P-PAL, uma vez que a

sobreposição com as fontes de alguns *corpora* jornalísticos e literários da Linguateca poderia reflectir-se numa sobrestimação dos valores de frequências a extrair. Contudo, a análise detalhada às fontes, títulos e anos de publicação das 186 obras literárias incluídas no CORLEX e das 217 obras literárias incluídas no Vercial, permitiu identificar 11 obras comuns, o que corresponde, num total de 403 obras, a uma sobreposição de apenas 2.73% do *corpus* literário total. No que se refere aos *corpora* jornalísticos, a análise às fontes dos 16,723 artigos integrados no sub*corpus* jornalístico do CORLEX revelou que 4,697 (28%) correspondem a publicações do *Jornal Público* dos anos 1997 e 1998. Esta situação poderá indiciar uma potencial sobreposição entre os artigos do *Jornal Público* integrados no CORLEX e no CETEMPúblico, muito embora o CETEMPúblico integre, como vimos, artigos de um período de tempo mais alargado (publicações de 1991 a 1998). Todavia, mesmo nesta situação, há que considerar que o impacto desta potencial sobreposição num *corpus* jornalístico de ≈200 milhões de palavras, como é o caso do CETEMPúblico, corresponderá a uma sobreposição de apenas 1.2%. Esta percentagem, à semelhança da do *corpus* literário, terá assim um impacto muito pouco significativo na sobrestimação dos valores de frequências a extrair no P-PAL. Por isso, e na impossibilidade de verificar estes textos manualmente, optámos pela integração do CORLEX, que contribui, no nosso entender, de forma significativa para o enriquecimento e diversificação dos géneros e modalidades linguísticas no P-PAL.

A Figura 1 apresenta a distribuição por género (técnico-científico e didáctico, literário, jornalístico e miscelânea) e modalidade discursiva (oral e escrito) dos *corpora* utilizados para indexação das entradas, frequências e categorias do P-PAL.

<INSERIR FIGURA 1>

O P-PAL integra essencialmente registos de língua escrita (226,552,040 palavras) e um pequeno sub*corpus* de língua falada (1,218,712 palavras). A maior proporção é

jornalística (94.5% do *corpus* total). Neste género, o CETEMPúblico é aquele que concorre com a percentagem mais significativa de ocorrências (89.1%), seguindo-se o CORLEX (4%), o DiaCLAVE (3.1%), o Avante! (3%) e o NaturaMinho (0.8%). O género literário representa 3.4% do *corpus* total. Neste género, a maior proporção deriva do Vercial, que concorre com 60% das ocorrências. O género técnico-científico e didáctico representa 1.6% do *corpus* total, contribuindo a porção do CORLEX de forma mais significativa para a sua composição (99.3%). O ECI-EE contribui com apenas 0.7% das ocorrências. Incorporámos ainda o género "miscelânea" do CORLEX que integra 575,962 ocorrências correspondentes a 0.3% do *corpus* total e escrito.

Da compilação destes oito *corpora* resultou assim um *corpus* total de 227,770,752 ocorrências provenientes de texto predominantemente escrito e jornalístico. Esta situação não é de admirar se considerarmos a natureza e a dimensão dos *corpora* integrados no P-PAL. Com efeito dos sete *corpora* da Linguateca, quatro são do género jornalístico, sendo que destes o CETEMPúblico é aquele que apresenta a maior dimensão de todos os *corpora* integrados no P-PAL. Apesar deste desequilíbrio na distribuição do género, consideramos que a inclusão de vários títulos de jornais provenientes de diferentes regiões do país e anos de publicação (1991-2002) podem concorrer para a obtenção de um léxico do PE mais diversificado no P-PAL e assim para aumentar a representatividade da língua.

Procedimento

A compilação de vários *corpora* para criação de um único léxico de frequências coloca grandes desafios ao tratamento da informação. Porque o léxico do P-PAL deriva de oito *corpora* pré-existentes assentes em sistemas de classificação morfo-sintáctica e de lematização distintos, tivemos de proceder, antes da extracção do léxico, a uma análise aprofundada aos sistemas de classificação morfo-sintáctica e de lematização adoptados em

cada um, com vista à normalização da terminologia e à criação de um sistema de classificação comum que rentabilizasse a informação original disponibilizada a partir de cada um deles.

A Tabela 1 apresenta os sistemas de classificação morfo-sintáctica adoptados nos *corpora* da Linguateca e no CORLEX, bem como o sistema adoptado no P-PAL. As ocorrências em cada *corpus* encontram-se convertidas numa escala logarítmica de base 10 (log_{10}).

<INSERIR TABELA 1>

Como se pode observar na Tabela 1, há um maior número de subcategorias nos *corpora* da Linguateca do que no CORLEX. Por exemplo, no CORLEX os nomes próprios não constituem entrada mas a Linguateca distingue entre nomes próprios e nomes próprios com designação comercial. Os pronomes estão classificados como pessoal, demonstrativo, indefinido, possessivo, interrogativo e relativo. Em contrapartida na Linguateca os pronomes são subcategoria das categorias determinante (DET) ou Especificador (SPEC), podendo pertencer às duas. Os artigos estão classificados como subcategoria da categoria principal DET, que não consta nas categorias do CORLEX, onde os artigos constituem categoria principal. Os *corpora* da Linguateca incluem ainda as categorias principais DET e SPEC, que integram as subcategorias artigo, pronome e adjectivo e pronome e adjectivo, respectivamente.

No que se refere à classificação de lemas e formas registam-se também diferenças. Nos *corpora* da Linguateca são lema os nomes no masculino e no feminino singular (“carcereiro” e “carcereira”, “imperador” e “imperatriz”), todas as palavras invariáveis (como as preposições, as conjunções e os advérbios, excepto as contracções e as locuções), os verbos no infinitivo impessoal (“estar”, “poder”, “fazer”) e os adjectivos no masculino singular (“novo”, “bom”), à excepção dos adjectivos com função de nome. Nestes casos são utilizados lemas diferentes para o feminino e o masculino (ex., o lema de “professores” é “professor” e o

de “professoras” é “professora”). Os pronomes pessoais têm como lema o pronome pessoal recto no masculino (ex., “eu” é lema de “me”, ”nós” lema de “nos”, “eles” lema de “lhe”, “lhes” ou “lhas”) e os pronomes possessivos, relativos, demonstrativos e interrogativos têm como lema o masculino singular (ex., “meu” é lema de “meus”, “minha” e “minhas”, “cujo” é lema de “cujos”, “cuja” e “cujos”). Constituem também lema as palavras compostas hifenizadas (ex., “aéreo-terrestre”) e não hifenizadas, sendo que, neste último caso, se assinalam com o símbolo “=” (ex., “*ad=hoc”*).

À semelhança dos *corpora* da Linguateca, no CORLEX são lema as palavras invariáveis (como as preposições, conjunções e advérbios, excepto as contracções e as locuções), os verbos no infinitivo impessoal, os nomes no masculino singular, embora os pronomes se apresentem tanto no masculino como no feminino singular. Quanto às palavras compostas (no sentido lato do termo, incluindo compostos morfológicos, compostos morfo-sintácticos e compostos sintácticos), o critério de lematização adoptado não é consistente. Por exemplo, “abelha-mãe” está inserida no lema “abelha”, mas “abelha-mestra” tem entrada própria na lista de lemas. Observa-se ainda que algumas palavras hifenizadas constituem lema de itens multilexicais não hifenizados (ex., “à-vontade” é lema das formas “à” e “vontade” e “abaixo-assinado” é lema das formas “abaixo” e “assinado*”*).

Inspirados pela proposta do Dicionário da Língua Portuguesa Contemporânea de Casteleiro (2001), adaptámos os sistemas de classificação morfo-sintáctica da Linguateca e do CORLEX numa nova classificação (cf. Tabela 1). O P-PAL contempla assim 10 categorias principais: nomes (N), determinantes (DET), pronomes (PRON), quantificadores (QUANT), adjectivos (ADJ), verbos (V), interjeições (INT), preposições (PREP), advérbios (ADV) e conjunções (CONJ). Os DET podem ser ainda classificados como demonstrativos, possessivos, indefinidos, relativos e interrogativos e os artigos como definidos ou indefinidos. Os PRON estão classificados como pessoais, demonstrativos, indefinidos, possessivos,

interrogativos e relativos e os QUANT como universais, existenciais, relativos e interrogativos, incluindo-se também nesta classe os numerais cardinais, ordinais, multiplicativos e fraccionários. O P-PAL contém ainda os ADV interrogativos e as CONJ subordinativas e coordenativas.

Atendendo aos diferentes critérios de lematização dos *corpora* da Linguateca e do CORLEX definimos, à semelhança da classificação morfo-sintáctica, um modelo específico de lematização para o P-PAL que optimizasse a informação oriunda de cada *corpus*. Assim, constituem lemas no P-PAL: (i) os verbos no infinitivo impessoal (ex., "abrir"); (ii) os nomes no masculino singular (ex., "gato", "padeiro"). Para os nomes de género fixo (masculino ou feminino) escolheu-se a forma singular (ex., "animal", "cobra"). Os nomes invariáveis quanto ao número (ex., "pires*'*") ou cuja flexão de género deriva de um radical distinto (ex., "homem/mulher", "cavalo/égua") constituem entrada própria; (iii) os adjectivos no masculino singular (ex., "bonito"). Para os adjectivos de género fixo é usada a forma singular (ex., "fácil"); (iv) os determinantes e os pronomes encontram-se no masculino e no feminino singular; (v) as classes invariáveis, advérbios, preposições, conjunções e interjeições; (vi) os numerais que formam uma unidade lexical. Por exemplo, "quinze" é considerado lema, mas "mil e duzentos" constitui uma unidade multilexical. A frequência destes itens multilexicais foi somada às frequências de cada um dos seus constituintes (i.e., aos lemas "mil", "e" e "duzentos"); (vii) as siglas e acrónimos considerados nomes comuns de acordo com Casteleiro (2001), excepto os que assumem função de nomes próprios, como o caso de algumas organizações (ex., "GNR", NATO), partidos e movimentos; e (viii) todos os vocábulos hifenizados que possuem entrada própria nos dicionários de referência (Casteleiro, 2001) e as palavras formadas por derivação prefixal cujo afixo altera o significado (ex., "anti-adiposo" *vs.* "adiposo"), o referente (ex., "auto-estrada" *vs.* "estrada") ou a classe do radical (ex., "além-fronteiras"$_{ADV}$ *vs.* "fronteira"$_{N}$). Foram ainda incluídos os estrangeirismos que

constituem entrada própria nos dicionários de referência. Contudo, é de assinalar que os estrangeirismos que apresentam uma ortografia não adaptada ao PE (ex., "*timing"*, "*briefing"*) foram incluídos como entrada mas excluídos da computação das restantes métricas do P-PAL, visto que a sua grafia não corresponde à grafia convencional do PE.

Não constituem entrada de lema no P-PAL as contracções, embora sejam contabilizadas na sua decomposição (ex., " dele" foi decomposto na preposição "de" e no pronome pessoal "ele" e a frequência atribuída a ambos os lemas), as unidades multilexicais disjuntas (i.e., não hifenizadas, onde se incluem as locuções, as expressões idiomáticas ou as colocações), embora as suas unidades constituintes tenham sido lematizadas e incluídas como entrada de lema e a frequência somada a cada um dos lemas reconstituídos; os nomes próprios, identificados a partir da informação disponibilizada pelo Portal do Cidadão, do Instituto dos Registos e do Notariado e na página do COMPARA disponível na Linguateca; e as palavras formadas por derivação prefixal cujas partículas não possuam existência própria na língua (ex., a partícula "recém" em "recém-chegado"), sendo que, nestes casos, se lematizou o radical ("chegar"), que assume o valor de frequência da forma composta original. Foram também excluídas abreviaturas (ex., "*vol*." ou "*art.*"), símbolos e grafias não convencionais (ex., "@", "S").

Da base das formas fazem parte todas as formas gráficas pertencentes às classes morfo-sintácticas adoptadas no P-PAL (cf. Tabela 1), incluindo palavras homónimas gramaticalmente distintas que, embora possuam grafia igual, pertencem a classes gramaticais diferentes (ex., "além"$_{N}$ *vs*. "além"$_{ADV}$). Neste grupo de palavras com grafia e fonética iguais a desambiguação fez-se a partir da categoria morfo-sintáctica, sendo contabilizadas como entradas distintas. As formas compostas hifenizadas foram também incluídas como entrada de acordo com a sua ocorrência natural no *corpus* e, ao contrário do procedimento adoptado para

os lemas, não foram decompostas nos seus itens constituintes. Fazem ainda parte da base de formas os estrangeirismos incluídos na base de lemas.

Por último submetemos os verbos flexionados e hifenizados com pronomes clíticos a um tratamento específico, uma vez que representam duas palavras numa forma composta. Nesse sentido, reconstituímos os verbos e atribuímos o valor de frequência original à forma verbal e ao pronome clítico correspondente, que também constitui entrada. Por exemplo, as formas verbais terminadas em "á" seguidas de pronome clítico (ex., "encestá-la") foram reconstituídas substituindo "á " por "ar" e suprimindo o pronome. As formas terminadas em "ávamo" (ex., "ajudávamo-nos") foram reconstituídas, substituindo *"ávamo"* por *"ávamos"* e suprimindo o pronome clítico. Às formas verbais terminadas em "ava" (ex., "perfilava-se") e em "avam" (ex., "preparavam-se"), não carecendo de reconstituição, subtraiu-se apenas o pronome clítico. Para as formas verbais seguidas dos pronomes "lo", "la", "los" ou "las" adoptámos um procedimento semelhante mas a frequência foi contabilizada nos pronomes pessoais "o", "a", "os" e "as", respectivamente.

Depois de definidos os sistemas de anotação morfo-sintáctica e de classificação de lemas e formas no P-PAL, procedemos a um conjunto de procedimentos de limpeza e de verificação automática e manual da informação. Atendendo ao elevado número de erros ortográficos e morfo-sintácticos existentes decorrentes de uma anotação automatizada e não revista nos *corpora* da Linguateca, para rentabilizar as verificações ao léxico extraído no P-PAL, implementámos um conjunto de procedimentos complementares. Numa primeira fase usámos o analisador morfológico JSpell (Simões & Almeida, 2001) para verificação automática da ortografia e da informação morfo-sintáctica das formas e lemas dos *corpora* da Linguateca, que depois cruzámos com a informação do *corpus* CORLEX, que foi verificado manualmente. Esta primeira verificação automatizada permitiu eliminar números e palavras com caracteres não convencionais, identificar palavras novas ou erros de ortografia e detectar

categorias não reconhecidas (e que, por isso, poderiam representar erros de anotação morfo-sintáctica). As entradas e etiquetas não constantes no JSpell ou no CORLEX foram verificadas manualmente e corrigidas se necessário. Estas verificações e correcções foram realizadas sequencialmente e iniciaram-se pelo CETEMPúblico que, pela sua dimensão, permitiu criar uma base de erros ortográficos e morfo-sintácticos comuns. O desenvolvimento desta base de erros foi essencial, acrescentando gradualmente informação de cada *corpus*. Cruzando essa base nos *corpora* seguintes, diminuímos progressivamente o volume de verificações manuais a realizar.

Os pares palavra/categoria não reconhecidos pelos procedimentos descritos acima foram verificados manualmente e re-etiquetados. Assim, por exemplo, no CETEMPúblico registaram-se ocorrências da entrada "bocado" anotada como V. Visto que a palavra apenas poderá ocorrer na língua com a função de N, alterámos a sua categoria. Registámos ainda ocorrências da palavra "sobre" como ADV. Ora, sabendo que na língua a palavra pode ocorrer como PREP e V e porque o elevado número de incidências inviabilizava a sua desambiguação contextual, optámos por atribuir, nestes casos, todas as categorias que a palavra pode assumir. Assim, se no P-PAL uma dada entrada estiver associada a mais do que uma categoria, deverá ter-se em conta que as ocorrências originais nos *corpora* podem corresponder a qualquer uma ou a todas as categorias a ela associadas. Um exemplo concreto de uma adaptação semelhante é a forma "se", que no CETEMPúblico apresenta valores de frequência para cinco categorias distintas e que foi integrada no P-PAL como uma única entrada associada às categorias CONJ subordinada e PRON pessoal.

Noutros casos a ambiguidade sintáctica gerou ambiguidade na lematização. As formas verbais "fora", "vendo" e "vimos", por exemplo, extraídas da lista de lemas e resultantes de erros de lematização das bases originais, podem corresponder aos lemas "ser/ir", "vender/vendar/ver" e "ver/vir", respectivamente. Na impossibilidade de verificação manual

destas ocorrências, optámos pela exclusão destas palavras da base de lemas, sob pena de sobrevalorizamos a frequência de algum desses lemas. Estas palavras constam assim como entradas na base de formas mas as suas frequências não foram associadas a nenhum lema.

Por último, cabe referir o tratamento às vacilações entre a hifenização e a disjunção (ex., "fim-de-semana" *vs.* "fim de semana") e à identificação e lematização dos compostos disjuntos. A primeira tarefa passou pela identificação de todas as palavras hifenizadas que constituem lemas nos *corpora* originais e na análise da sua lematização. Os itens multilexicais disjuntos foram decompostos em unidades separadas, sendo a frequência atribuída a cada um dos seus constituintes. Por exemplo, a frequência de "água de colónia" (forma ou lema), foi adicionada às frequências das entradas "água", "de" e "colónia". Contudo, nas situações em que um dos itens constituintes não ocorre isoladamente na língua (ex., "verdiano" em "cabo verdiano", "iorquino" em "nova iorquino" e "versa" em "vice versa"), essas palavras foram eliminadas da base. As unidades multilexicais hifenizadas foram incluídas como entrada própria, pelo que "água-de-colónia" constitui entrada tanto na base de formas como de lemas no P-PAL. Como resultado deste processo de verificações foram eliminadas dos oito *corpora* originais 4,422,753 ocorrências de formas e 1,402,546 ocorrências de lemas.

Resultados

O conjunto de procedimentos desenvolvidos na análise, tratamento e compilação dos oito *corpora* que integram o P-PAL permitiu obter um léxico constituído por 208,642 formas e 52,404 lemas, cuja distribuição por extensão de palavra (número de letras) se apresenta na Figura 2.

<INSERIR FIGURA 2>

Como podemos observar na Figura 2, o P-PAL inclui, na base de formas, palavras que variam de 1 a 31 letras e, na base de lemas, palavras que variam de 1 a 24 letras. A maioria

das palavras no P-PAL apresenta entre 7 a 11 letras que constituem 63.5% e 61.5% do léxico de formas e lemas, respectivamente. A extensão média das palavras no P-PAL situa-se na base de formas em 9.9 letras (*DP* = 2.97) e na base de lemas em 9.3 letras (*DP* = 2.96).

A Figura 3 apresenta a distribuição acumulada das frequências lexicais (por milhão de ocorrências) do P-PAL por extensão de palavra (número de letras) na base de formas e lemas.

<INSERIR FIGURA 3>

A análise à distribuição das frequências acumuladas revela uma distribuição tipo *Poisson*. Assim, à medida que avançamos na extensão de palavras (i.e., no número de letras que as integram) a probabilidade da sua ocorrência vai decrescendo de uma forma quase linear tanto em formas como em lemas, situando-se o ponto de corte em ambos casos em torno das 5 letras. A partir desse valor verifica-se uma quebra significativa nos valores de frequências acumuladas. Cabe no entanto assinalar que mais de 50% das frequências lexicais ocorrem em palavras com três ou menos letras no léxico de lemas (53.5%) e em palavras com quatro ou menos letras no léxico de formas (56.3%).

Com efeito, como se observa na Figura 3, as palavras de uma letra constituem as mais frequentes do léxico de lemas (entre as quais as palavras funcionais “a”, “e” e “o” com uma frequência por milhão de palavras de 88,046.59, 84,061.31 e 80,466.16, respectivamente). No léxico de formas as palavras de duas letras integram o conjunto das palavras mais frequentes, entre as quais se encontram as palavras funcionais “de” e “em” (com uma frequência por milhão de 46,474.75 e 12,561.91, respectivamente). Seguem-se, de forma muito aproximada, as palavras de uma letra (e que incluem, à semelhança do observado para os lemas, as palavras funcionais “a”, “a” e “o”, ainda que com uma distribuição de frequências distinta – 39,164.26, 87,551.52 e 30,020.17, respectivamente – às quais acresce a contracção “à” com 5,060.34 ocorrências e a forma verbal “é” com 7,391.74 ocorrências). As formas do P-PAL apresentam uma frequência lexical por milhão de palavras que varia entre 0 (palavras que

ocorrem apenas uma vez no *corpus* – 47,154 palavras) e 87,551.52 por milhão de ocorrências, com uma frequência média de 4.69 (*DP* = 272.18). A palavra mais frequente corresponde à palavra funcional "e". Na base de lemas, a frequência varia entre 0 (5,246 palavras) e 89,567.61por milhão de ocorrências, com uma frequência média de 18.93 (*DP* = 788.44). O lema mais frequente corresponde à palavra funcional "de".

Porque o P-PAL resulta, como vimos, da compilação de oito *corpora* de diferentes tipos (escrito e oral) e géneros linguísticos (jornalístico, literário, técnico etc.), procedemos a uma análise de correlação produto-momento *Pearson* entre a medida de frequência lexical obtida no P-PAL e as obtidas em cada um dos oito *corpora* que lhe deram origem. A Tabela 2 apresenta as correlações obtidas na base de formas (porção superior da tabela a cinzento) e de lemas (porção inferior da tabela).

<INSERIR TABELA 2>

Como se observa na Tabela 2, as correlações entre a medida de frequência do P-PAL e as restantes medidas de frequência oriundas de cada *corpus* são positivas e estatisticamente significativas tanto na base de lemas como na de formas, situando-se acima de 0.80 (excepção feita à correlação observada entre o *corpus* Museu Pessoa e Natura/Minho no caso das formas – r =0.75). Cabe no entanto assinalar a existência de correlações mais elevadas entre a medida de frequência do P-PAL e o *corpus* jornalístico CETEMPúblico tanto na base de formas como na de lemas (r = 0.99), o que não é de estranhar se considerarmos o peso que esse *corpus* representa no P-PAL (cf. Figura 1). Seguem-se, na base de formas, o *corpus* jornalístico Avante! e o literário Vercial, ambos com uma correlação situada nos 0.90. Na base de lemas, o CORLEX assume-se como o segundo *corpus* mais associado à medida de frequência do P-PAL (r = 0.95). As correlações menos elevadas (ainda que mesmo assim situadas num intervalo de elevada correlação) observam-se entre a medida de frequência do P-PAL e a do *corpus* jornalístico Natura/Minho, tanto na base de formas (r = 0.83) como na de lemas,

embora neste último caso este *corpus* se associe ao ECI-EE (apresentando em ambos casos uma correlação de 0.87). De assinalar ainda que, sendo um *corpus* de linguagem essencialmente escrita, o P-PAL apresenta elevada correlação com o *corpus* oral Museu Pessoa, com valores de correlação de 0.84 na base de formas e 0.85 na base de lemas.

Discussão

Neste trabalho apresentámos os procedimentos de compilação, análise e tratamento de oito *corpora* do PE contemporâneo de livre acesso e etiquetados morfossintaticamente (sete disponibilizados pela Linguateca e um pelo CLUL), com vista à criação de um *corpus* de grandes dimensões e diversificado na sua composição interna para a extracção de uma nova medida de frequência lexical disponibilizada a partir da aplicação P-PAL (http://p-pal.di.uminho.pt/about/tools). Os procedimentos de compilação, análise e tratamento da informação oriunda dos diferentes *corpora* implicaram uma reclassificação morfo-sintáctica e a adopção de critérios de lematização que permitissem a criação de um sistema de classificação comum e rentabilizassem a informação original disponibilizada a partir de cada um deles. Por este procedimento, e inspirados na proposta do Dicionário da Língua Portuguesa Contemporânea de Casteleiro (2001), o P-PAL contempla 10 categorias morfossintácticas principais distribuídas em duas bases lexicais distintas compostas por 52,404 palavras lematizadas e 208,642 palavras não lematizadas (formas) do PE contemporâneo.

De assinalar o tratamento dado às contracções, às unidades multilexicais disjuntas (i.e., não hifenizadas, onde se incluem as locuções, as expressões idiomáticas ou as colocações), às palavras formadas por derivação prefixal (cujas partículas não possuam existência própria na língua), e aos verbos flexionados e hifenizados com pronomes clíticos que, ocorrendo na língua, não constituem, no P-PAL, uma unidade lexical única. Nestes casos

procedemos, como vimos, à decomposição da palavra ou da unidade multilexical, nos seus elementos constituintes e à atribuição do valor de frequência original a cada um dos lemas ou formas reconstituídos. Consideramos que este procedimento, ainda que mais dispendioso do ponto de vista do tratamento da informação, poderá constituir um elemento importante na fiabilidade da medida de frequência obtida. Com efeito, a exposição a esse tipo de palavras, comuns na língua portuguesa, não deverá ser negligenciada neste tipo de medida, sob pena de se subestimar a sua ocorrência e de se introduzir erro adicional numa medida que procura, como vimos, reflectir o uso efectivo que os falantes fazem da língua.

Estes procedimentos, à semelhança das tarefas de limpeza e verificação da informação, permitiram obter um *corpus* de grandes dimensões (mais de 227 milhões de palavras) o que, atendendo às recomendações mais recentes da literatura (ex., Brysbaert & New, 2009; Brysbaert *et al*., 2011; Lee, 2003; Sardinha, 2004; Sinclair, 2005) poderá, desde logo, concorrer para a qualidade da medida de frequência lexical aqui extraída. Em todo o caso, à semelhança de vários projectos internacionais (ex., *English Lexicon Project* – ver Balota *et al*., 2007), estudos futuros deverão comprovar a qualidade desta medida a partir, por exemplo, da recolha de tempos de reconhecimento e/ou nomeação de um vasto conjunto de palavras com vista à determinação do seu poder preditivo. Estudos deste tipo são tanto mais relevantes quanto a investigação internacional mais recente comprova, como vimos, que a frequência de uso de palavras se assume como a variável mais potente na explicação do desempenho linguístico dos sujeitos (ex., Balota *et al.*, 2007; Brysbaert & Cortese, 2011; Brysbaert *et al*., 2011; Ferrand *et al.*, 2010; Keuleers *et al.*, 2010; Thompson & Desroches, 2009).

De assinalar também que, embora no *corpus* do P-PAL predominem registos de linguagem escrita, o que poderia colocar em causa a representatividade da língua que pretendíamos obter com a diversidade de géneros e tipos discursivos integrados, a análise de

correlação desenvolvida, tomando as frequências de cada um dos oito *corpora*, revela no entanto que a medida de frequência do P-PAL apresenta correlação elevada não só com a medida de frequência de todos os *corpora* escritos que lhe deram origem, mas também com o *corpus* oral Museu Pessoa (situando-se acima de 0.80 em ambos os casos). Estes valores parecem evidenciar, à semelhança do observado noutras línguas (ex., Alonso, Fernandez & Díez, 2011; Pastizzo & Carbonne, 2007), que as frequências lexicais computadas para a linguagem escrita no PE poderão ser tomadas como um bom indicador das frequências obtidas a partir da linguagem oral, e que o facto de o P-PAL integrar essencialmente informação oriunda de registos de linguagem escrita (e, dentro destes, de tipo jornalístico), poderá não constituir em si mesmo uma limitação à sua validade. Aliás, a inclusão no P-PAL de outros *corpora* permitiu enriquecer fortemente a sua diversidade linguística, dado que do *corpus* oral Museu Pessoa foram incorporadas no P-PAL apenas 14,259 formas e 6,934 lemas o que constitui somente 6.8% e 13.2% do léxico total respectivo. A inclusão dos restantes *corpora* escritos permitiu assim contribuir de forma significativa para a diversidade lexical do P-PAL e para a obtenção de um léxico representativo do PE contemporâneo, tal como se pretendia.

A análise à distribuição das entradas lexicais do P-PAL (formas e lemas) por extensão de palavra permitiu verificar também que, comparativamente a outras línguas, o PE apresenta, em média, palavras de maior extensão (cf. Hatzigeorgiu, Mikros, & Carayannis, 2001; Riedemann, 1996; Ziegler, 2000). Esta situação reflecte de algum modo o facto de o Português ser uma língua sintética, morfologicamente rica, na qual novas palavras podem ser formadas mediante a junção de morfemas já existentes por prefixação e/ou sufixação, como em “en-trincheira-mento” (derivação) ou "cant-á-va-mos" (flexão), ou mediante a junção de palavras ou radicais (composição), como “malmequer”. A este número há ainda a acrescentar a integração de palavras compostas hifenizadas (tanto compostos morfológicos – ex., “luso-

brasileiro"-; compostos morfossintáticos – ex., "surdo-mudo"- como conjuntos ou encontros ocasionais – ex., "integracionistas-centralizadoras") que no caso dos lemas constituem 1,770 entradas (3.4% do léxico total) e no caso das formas 18,911 entradas (9.1% do léxico total). A inclusão destas palavras faz incrementar a extensão média das palavras no P-PAL, de tal forma que se excluídas, a extensão das palavras oscilaria no caso das formas entre um mínimo de 1 e um máximo de 24 letras, com uma média de 9.6 letras (*DP* = 2.65) e no caso dos lemas entre um mínimo de 1 e um máximo de 22, com uma média de 9.2 letras (*DP* = 2.94).

Embora a literatura sobre os efeitos de extensão no reconhecimento visual de palavras seja inconsistente (ver New, Ferrand, Pallier, & Brysbaert, 2006 para uma revisão), os estudos realizados até ao momento foram maioritariamente conduzidos ora em línguas opacas (ex., inglês), ora em línguas transparentes (ex., espanhol), deixando o que se passa em línguas semi-transparentes, como o PE, por esclarecer. Além disso, a esmagadora maioria desses estudos recorreram a palavras monossilábicas de pequena extensão pelo que se questiona até que ponto não só os dados obtidos em línguas opacas ou transparentes se podem generalizar a outras línguas, como o facto de os resultados obtidos no mesmo idioma para palavras de pequena extensão poderem não ser generalizáveis para palavras de maiores extensões (Soares *et al.*, 2012). Nova investigação deverá pois ser desenvolvida para testar esses efeitos.

A análise à distribuição de frequências acumuladas revelou, como esperado, que à medida que a extensão de palavras aumenta, a probabilidade da sua ocorrência vai decrescendo de uma forma quase linear. À semelhança do observado noutras línguas (ver Grotjahn, & Altmann, 1993; Sigurd, Eeg-Olofsson, & van de Weijer, 2004; Wimmer & Altmann, 1996), esta relação comprova também no PE a lei de *Zipf*, segundo a qual as palavras mais frequentemente usadas numa língua são aquelas que requerem menos esforço no seu uso/utilização. No P-PAL mais de 50% das frequências lexicais ocorrem em palavras de três ou menos letras no léxico de lemas e de quatro ou menos letras no léxico de formas.

Dentro dessas, as palavras funcionais assumem os valores de frequência lexical mais elevados. Por último, resta assinalar que o P-PAL disponibiliza ainda um conjunto de outras medidas como o grau de similitude ortográfica e fonológica entre palavras (i.e., medidas de vizinhança). Estas medidas assumem elevada relevância na literatura, dada a constatação empírica de que o processamento de uma dada palavra conduz à activação automática de outras palavras similares, o que consequentemente afecta o seu acesso lexical (ver Andrews, 1997). As medidas de similitude ortográfica do P-PAL incluem, tanto na base de formas como de lemas, a medida *standard* de densidade de vizinhança de Coltheart, Davelaar, Jonasson e Besner (1977), que reflecte o número de palavras existentes no léxico que diferem da palavra alvo pela substituição de uma letra mantendo as restantes constantes nas mesmas posições (ex., a forma "alma" tem como vizinhos "alba", "alça", "alfa", "alga", "alia", "almo", "alta", "alva", "arma", "asma" e "alua", apresentando assim um valor de $N = 11$). Estas medidas contemplam ainda a distribuição das frequências desses vizinhos (ex., a média de frequência da vizinhança de "alma" é de 14.51, sendo que "alta" se releva o vizinho com a frequência mais elevada com 111.32 ocorrências por milhão de palavras contra os 57.63 da palavra alvo "alma"). O P-PAL inclui ainda medidas de densidade e frequência dos vizinhos gerados por adição (i.e., junção de uma letra à palavra alvo – ex., "alma" possui "calma" e "palma" como vizinhos por adição, cujas frequências de ocorrência são inferiores à de "alma" – 25.60 e 1.20 respectivamente), subtracção (i.e., eliminação de uma letra à palavra alvo – ex., "alma" apresenta "ala" e "ama" como vizinhos por subtracção, sendo que estes apresentam também frequências de ocorrência inferiores à de "alma" – 16.09 e 9.39 respectivamente) e transposição de letras (i.e., alteração da posição relativa de alguma das letras da palavra alvo – ex., "alma" apresenta "lama" e "alam" como vizinhos por transposição, que apresentam também frequências lexicais inferiores a "alma" – 10.67 e 0.27 ocorrências por milhão de palavras, respectivamente). Estas medidas de similitude ortográfica assumem também elevada

importância com contexto da investigação mais recente (ver Charles-Luce & Luce, 1990; Davis & Taft, 2005 Perea & Lupker, 2004). À semelhança das restantes línguas, as palavras que integram o P-PAL variam nessas métricas (ver Soares *et al*., 2011, 2012), pelo que os investigadores deverão atender a essas características no controlo e/ou manipulação dos estímulos na condução das suas investigações.

Conclusão

Em conclusão podemos afirmar que o PE dispõe na actualidade de um novo léxico de frequências que, partindo da rentabilização de *corpora* já existentes, permitiu a constituição de um *corpus* de grandes dimensões (mais de 227 milhões de palavras) e diversificado na sua composição interna. Ele inclui, ainda que de forma não equitativa, registos da linguagem oral e escrita oriundos dos mais variados géneros, desde o jornalístico, o literário e o técnico-científico ao didáctico, o que no nosso entender contribui de forma significativa para o enriquecimento da variedade lexical do seu *corpus* e, consequentemente, para a representação da língua e a afirmação da validade das medidas de frequência extraídas a partir dele.

Pelo potencial que oferece à investigação, ao disponibilizar numa aplicação informática amigável de acesso gratuito não só um novo índice de frequência lexical mas todo um conjunto de outros índices sobre as palavras do PE não disponíveis até então (ex., índices de similitude ortográfica, fonológica, fonográfica, silábica) consideramos que o P-PAL se assume como uma ferramenta sem par, de valor inestimável à promoção e à internacionalização da investigação em Portugal.

Agradecimentos

Agradecemos à FCT (Fundação para a Ciência e a Tecnologia), ao QREN (Quadro de Referência Estratégica Nacional) e ao COMPETE (Programa Operacional Factores de Competitividade), integrado no Fundo Europeu de Desenvolvimento Regional (FEDER), o financiamento deste projecto (PTDC/PSI-PCO/104679/2008).

Agradecemos à Linguateca, em particular à Doutora Diana Santos, pela colaboração na disponibilização do *corpus* Vercial por séculos.

Agradecemos ainda à Doutora Maria Fernanda Bacelar do Nascimento e ao CLUL pelo envio das fontes do sub*corpus* literário do CORLEX.

Referências

Alonso, M. A., Fernandez, A., & Díez, E. (2011). Oral frequency norms for 67,979 Spanish words. *Behavior Research Methods, 43,* 449-458.

Andrews, S. (1997). The role of orthographic similarity in lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin and Review, 4,* 439-461.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database* (Release 2) [CD-ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bacelar do Nascimento, M. F., Pereira, L. A. S., & Saramago, J. (2000). Portuguese Corpora at CLUL. In M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, & G. Steinhaouer (Eds.), *Second International Conference on Language Resources and Evaluation – Proceedings,* Volume II (pp. 1603-1607). Athens: European Language Resources Association.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133,* 283-316.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B. … Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445-459.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Access and/or production? *Journal of Memory and Language, 24*, 89-106.

Boudelaa, S., & Marslen-Wilson, W. D. (2010). Aralex: A lexical database for modern standard arabic. *Behavior Research Methods, 42,* 481–487.

Brysbaert, M., & Cortese, M. J. (2011). Do the effects of subjective frequency and age of acquisition survive better word frequency norms? *Quarterly Journal of Experimental Psychology*, *64*, 545-559.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, Instruments & Computers*, *41*, 977-990.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412-424.

Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kŭcera and Francis. *Behavior Research Methods, Instruments, & Computers*, *30*, 272-277.

Carroll, J. B., Davies, P., & Richman, B. (Eds.) (1971). *The American Heritage word-frequency book*. Boston: Houghton Mifflin.

Casteleiro, J. M. (coord.) (2001). *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. Lisboa: Academia das Ciências de Lisboa/Editorial Verbo.

Cattell, J. M. (1886). The time it takes to see and name objects. *Mind*, *11*, 63-65.

Charles-Luce, J. & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language*, *17*, 205–15.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Coltheart, M., Davelaar, E., Jonasson, J. F., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention & Performance VI* (pp. 535-555). Hillsdale, NJ: Erlbaum.

Content, A., Mousty, P., & Radeau, M. (1990). BRULEX. Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, *90*, 551-566.

Costa, L., Santos, D., & Cardoso, N. (Eds.) (2008). *Perspectivas sobre a Linguateca. Actas do encontro Linguateca: 10 anos*. Linguateca.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*, 317-367

Davis, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, 37(1),* 65-70.

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, 37(4),* 665-671.

Davis, C. J., & Taft, M. (2005). More words in the neighborhood: Interference in lexical decision due to deletion neighbors. *Psychonomic Bulletin & Review, 12*, 904-910.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A. … Pallier, C. (2010). The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods, 42(2),* 488-496.

Forster, K. I., & Hector, J. (2002). Cascaded versus noncascaded models of lexical and semantic processing: The turple effect. *Memory & Cognition, 30(7),* 1106-1117.

Forster, K. I., & Shen, D. (1996). No enemies in the neighborhood: Absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning Memory and Cognition, 22(3),* 696-713.

Forster, K., & Chambers, S. (1973). Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior, 12,* 627-635.

Grotjahn, R., & Altmann, G. (1993). Modelling the distribution of word length. In R. Köhler, & B. B. Rieger (Eds.), *Contributions to quantitative linguistics* (pp. 141 –153). Dordrecht: Kluwer.

Hatzigeorgiu, N., Mikros, G., & Carayannis, G. (2001). Word length, word frequencies and Zipf's Law in the Greek language. *Journal of Quantitative Linguistics*, *8*, 175-185.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Frontiers in Psychology*, *1*.

Ktori, M., van Heuven, W. J. B, & Pitchford, N. J. (2008). GreekLex: A lexical database of Modern Greek. *Behavior Research Methods, 40(3),* 773-783.

Kučera, M., & Francis,W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Lee, C. J. (2003). Evidence-based selection of word frequency lists. *Journal of Speech-Language Pathology and Audiology*, *27(3),* 172-175.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instrumentation, and Computers, 28,* 203-208.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team … Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, *331*, 176–182.

Murray, W. S., & Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review, 111,* 721-756.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Re-examining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin & Review*, 13(1), 45-52.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36(3)*, 516-524.

Pastizzo, M. J., & Carbone, R. F. (2007). Spoken word frequency counts based on 1.6 million words in American English. *Behavior Research Methods, 39*, 1025-1028

Perea, M., & Lupker, S. J. (2004). Can CANISO activate CASINO? Transposed-letter similarity effects with nonadjacent letter positions. *Journal of Memory & Language*, *51*, 231-246.

Português Fundamental (1984). *Vocabulário e Gramática* (tomo 1). Lisboa: INIC.

Riedemann, H. (1996). Word-length distribution in English press texts. *Journal of Quantitative Linguistics*, *3*(*3*), 265-271.

Sardinha, B. T. (2004). *Lingüística de corpus*. Barueri: Manole.

Sebastián-Gallés, N., Martí, M. A., Cuetos, F., & Carreiras, M. (2000). *LEXESP: Léxico informatizado del español*. Barcelona: Adiciones de la Universitat de Barcelona.

Simões, A. M., & Almeida, J. J. (2001). Jspell: Um módulo de análise morfológica para uso em Processamento de Linguagem Natural. In A. Gonçalves, & C.N. Correia (Orgs), *Actas do Encontro Nacional da Associação Portuguesa de Linguística* (pp. 485-495). Lisboa: Associação Portuguesa de Linguística.

Sigurd, B., Eeg-Olofsson, M., & van de Weijer, J. (2004). Word length, sentence length and frequency - Zipf revisited. *Studia Linguistica*, *59*(*1*), 37-52

Sinclair, J. (2005). Corpus and text: Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (1-16). Oxford: Oxbow Books.

Soares, A. P., Comesaña, M., Iriarte, A., Almeida, J. J., Simões, A., Costa, A. … Machado, J. (2010). P-PAL: Uma base lexical com índices psicolinguísticos do Português Europeu. *Linguamática, 2(3),* 67-72.

Soares, A. P., Comesaña, M., Iriarte, A., Almeida, J. J., Simões, A., Costa, A. … Machado, J. (2011). Procura-PALavras (P-PAL): A web application for a new European Portuguese lexical database. Poster apresentado no *17th meeting of European Society of Cognitive Psychology* (ESCOP). Espanha: San Sebastián.

Soares, A. P., Costa, A., Machado, J., Iriarte, A., Simões, A., Almeida, J. J., … Comesaña, M. (2012). Procura-PALavras (P-PAL): Uma aplicação web para uma nova base lexical do português europeu. Poster apresentado no *7º Encontro da Associação Portuguesa de Psicologia Experimental* (APPE). Lisboa: Universidade de Lisboa.

Soares, A. P., Nascimento, A., Silva, A. M., Costa, A., Machado, J., Comesaña, M. … Perea, M. (2012). Efeitos de extensão e frequência lexical no reconhecimento visual de palavras do Português Europeu. Poster apresentado no *III Seminário de Investigação em Psicologia da Universidade do Minho (SIPUM)*. Braga, Universidade do Minho, Portugal.

Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods, 41(2),* 452-71.

Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Teachers College, Columbia University, 1944.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, NY: Touchstone Applied Science.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory & Language, 47,* 1-29.

Ziegler, A. (2000). Word length in romance languages: A complemental contribution. *Journal of Quantitative Linguistics*, *7*(*1*), 65-68.

Wimmer, G., & Altmann, G. (1996). The theory of word length: Some results and generalizations. *Glottometrika, 15*, 112 – 133

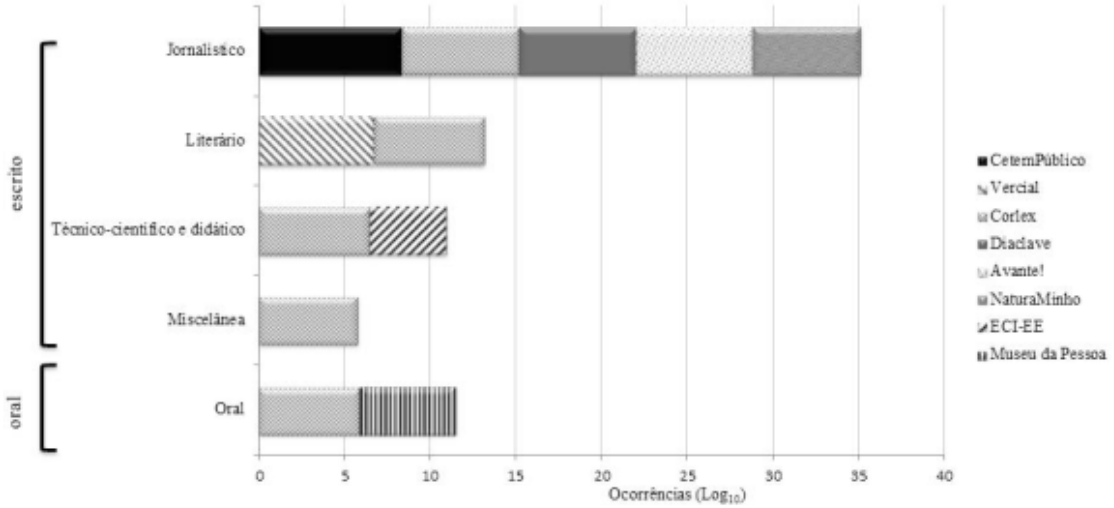

Figura 1: Distribuição dos *corpora* do P-PAL por género e tipo linguísticos.

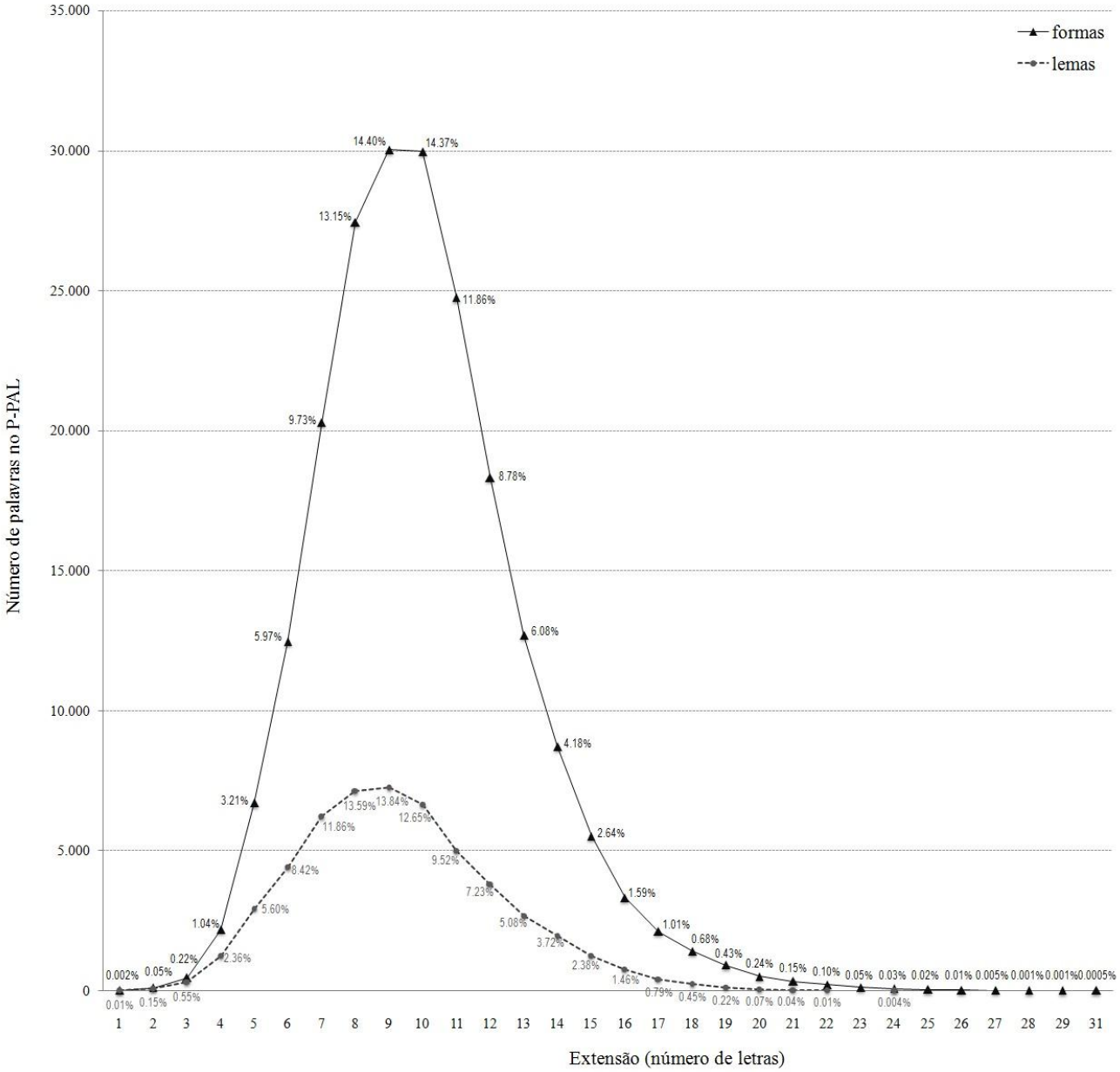

Figura 2: Distribuição das 208,642 formas e 52,404 lemas do P-PAL por extensão de palavra (número de letras).

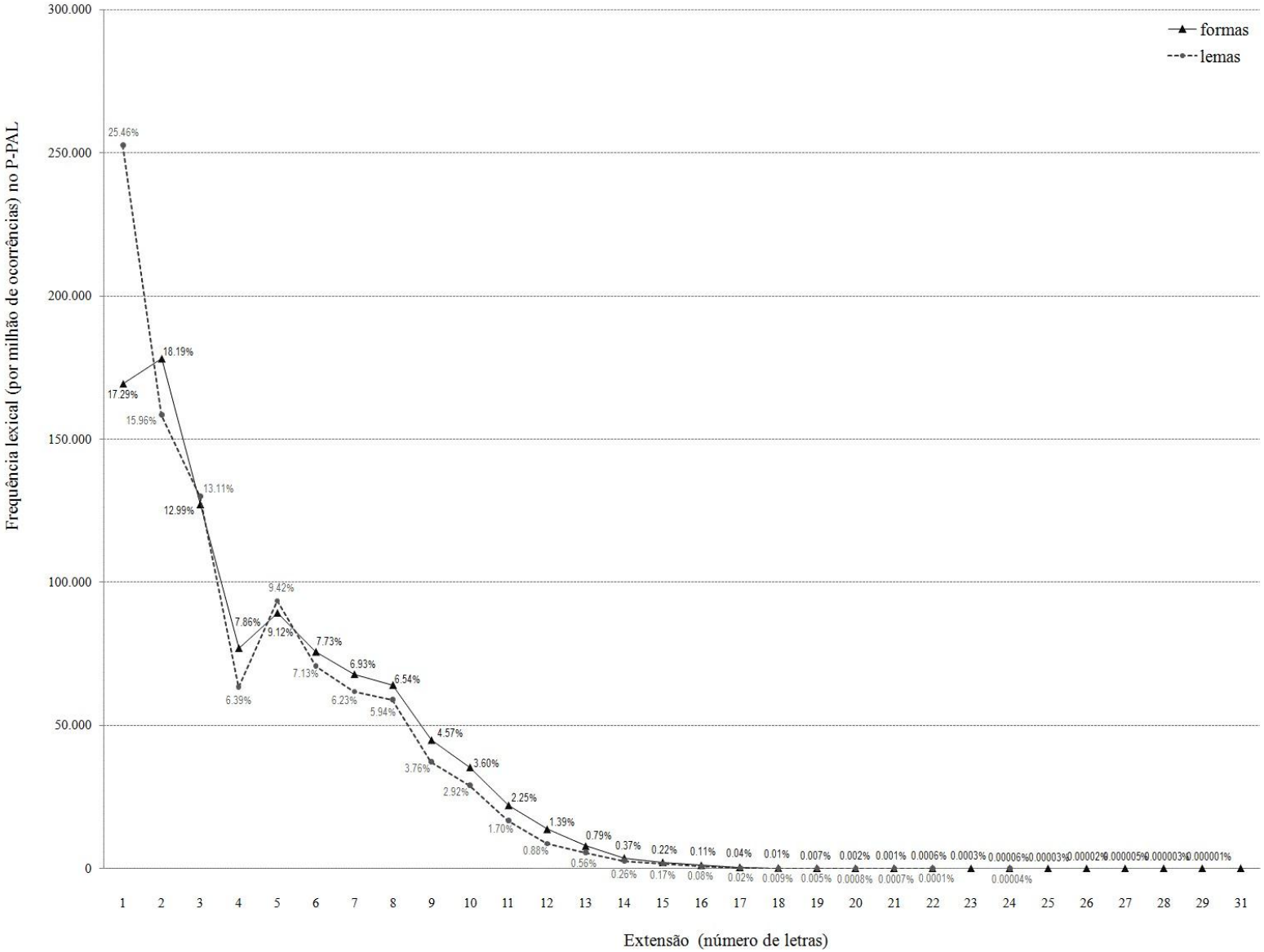

Figura 3: Distribuição das frequências acumuladas (por milhão de ocorrências) das 208,642 formas e 52,404 lemas do P-PAL por extensão de palavra (número de letras).

Tabela 1: Sistemas de classificação morfo-sintáctica adoptados nos *corpora* da Linguateca, no CORLEX e no P-PAL.

Corpora da Linguateca			CORLEX		P-PAL		
Nome (N) Nome próprio (PROP) e com designação comercial "&" (KC)			Nome (N)		Nome (N)		
Determinante (DET) (Artigos, Pronomes, Adjectivos)	Artigo	Definido (artd) Indefinido (arti)	Artigo (T)	Definido (Td) Indefinido (Ti)	Determinante (DET)	Artigo (ART	Definido (DEF) Indefinido (IND)
	Relativo (rel) Interrogativo (interr)					Demonstrativo (DEM) Possessivo (POSS) Indefinido (IND) Relativo (REL) Interrogativo (INT)	
Pronome pessoal (PERS)			Pronome (P)	Pessoal (Pp) Demonstrativo (Pd) Indefinido (Pi) Possessivo (Po) Interrogativo (Pt) Relativo (Pr)	Pronome (PRON)	Pessoal (PESS) Demonstrativo (DEM) Indefinido (IND) Possessivo (POSS) Interrogativo (INT) Relativo (REL)	
Especificador (SPEC) (Pronomes, Adjectivos)	Demonstrativo (dem) Possessivo (poss) Interrogativo (interr) Relativo (rel)						
Numeral (NUM)	Cardinal (card) Ordinal (ord) Fraccionário (fract)		Numeral (M)		Quantificador (QUANT)	Numeral (NUM)	Cardinal (CARD) Ordinal (ORD) Multiplicativo (MULT) Fraccionário (FRAC)
						Universal (UNI) Existencial (EXIS) Relativo (REL) Interrogativo (INT)	
Adjectivo (ADJ)			Adjectivo (A)		Adjectivo (ADJ)		
Verbo (V)	Principal	Verbo intransitivo (vi) Verbo transitivo (vt) Verbo transitivo directo (vtd)	Verbo (V)		Verbo (V)		
	Copulativo	vK e vtK					
Advérbio (ADV)			Advérbio (R)		Preposição (PREP)		
					Advérbio (ADV)	Interrogativo (INTR)	
Conjunção	Subordinativa (KS) Coordenada (KC)		Conjunção (C)		Conjunção (CONJ)	Subordinativa (SUB)	
						Coordenada (COOR)	
Interjeição (IN) Contracção (CONT)			Interjeição (I) Contracção (+)		Interjeição (INT)		
Divisão de itens multilexicais (MWE)	1º elemento da contracção (sam-) 2º elemento da contracção (-sam)		Elemento de Locução (L)				

Tabela 2: Correlações lineares (*Pearson*) entre as medidas de frequência por milhão de palavras obtidas no P-PAL, CETEMPúblico, Avante!, DiaCLAVE, Natura/Minho, ECI-EE, Museu da Pessoa, Vercial e CORLEX para a base de formas (porção de cima) e de lemas (porção de baixo).

	P-PAL	CETEMPúblico	Avante!	DiaCLAVE	Natura/Minho	ECI-EE	Museu Pessoa	Vercial	CORLEX
					formas				
P-PAL	-	0.99**	0.90**	0.87**	0.83**	0.87**	0.84**	0.90**	0.89**
CETEMPúblico	0.99**	-	0.89**	0.85**	0.82**	0.86**	0.82**	0.87**	0.87**
Avante!	0.89**	0.87**	-	0.99**	0.90**	0.95**	0.88**	0.95**	0.99**
DiaCLAVE	0.88**	0.86**	0.99**	-	0.90**	0.95**	0.88**	0.95**	0.99**
Natura/Minho	0.87**	0.86**	0.96**	0.97**	-	0.89**	0.75**	0.80**	0.88**
ECI-EE	0.87**	0.86**	0.97**	0.98**	0.98**	-	0.81**	0.88**	0.95**
Museu Pessoa	0.85**	0.84**	0.90**	0.90**	0.86**	0.85**	-	0.95**	0.90**
Vercial	0.90**	0.89**	0.98**	0.98**	0.93**	0.94**	0.94**	-	0.96**
CORLEX	0.95**	0.93**	0.91**	0.91**	0.90**	0.89**	0.85**	0.91**	-
					lemas				

** $p < 0.001$