P-PAL: Uma base lexical com índices psicolinguísticos do Português Europeu

Ana Paula Soares[1], Montserrat Comesaña[1], Álvaro Iriarte[2], José João de Almeida[3], Alberto Simões[3], Ana Costa[4], Patrícia Cunha França[4], João Machado[4]

[1]Escola de Psicologia, Universidade do Minho
[2]Instituto de Letras e Ciências Humanas, Universidade do Minho
[3]Departamento de Informática, Universidade do Minho
[4]Centro de Investigação em Psicologia, Universidade do Minho

{asoares,mvila}@psi.uminho.pt, alvaro@ilch.uminho.pt
{jj,ambs}@di.uminho.pt, {ana.costa,patfranca,joaoffm}@psi.uminho.pt

Resumo

Neste trabalho apresentamos o projecto Procura-PALvras (P-PAL) cujo principal objectivo é desenvolver uma ferramenta electrónica que disponibilize informação sobre índices psicolinguísticos objectivos e subjectivos de palavras do Português Europeu (PE). O P-PAL será disponibilizado gratuitamente à comunidade científica num formato amigável a partir de um sítio na Internet a construir para o efeito. Ao utilizar o P-PAL, o investigador poderá fazer uma utilização personalizada do programa ao seleccionar, da ampla variedade de análises oferecidas, os índices que se adequam aos propósitos da sua investigação e numa dupla funcionalidade de utilização: pedir ao programa para analisar listas de palavras previamente constituídas nos índices considerados relevantes para a investigação ou para obter listas de palavras que obedeçam aos parâmetros definidos. O P-PAL assume-se assim como uma ferramenta fundamental à promoção e internacionalização da investigação em Portugal.

1 Introdução

A importância da existência de bases lexicais informatizadas que apoiem de forma efectiva a investigação nas áreas da Psicologia Cognitiva, das Neurociências, da Linguística ou do Processamento de Linguagem Natural (PLN) é, na actualidade, um dado 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, em si mesmas, um estímulo complexo, que reúnem um conjunto de propriedades ou atributos cujo controlo e/ou manipulação se revelam fundamentais ao desenvolvimento profícuo de estudos nesses domínios, a investigação nacional e internacional já não se compadece mais com a inexistência deste tipo de ferramentas.

Refira-se, a título de exemplo, a sua utilidade, nas áreas mais experimentais da Psicolinguística ou das Neurociências, onde o seu apoio à selecção de estímulos (palavras) se revela essencial. Entre as características que se desejam ver devidamente manipuladas ou controladas, encontram-se tanto propriedades mais objectivas, que podem ser determinadas directamente pela análise da própria palavra (p. ex. extensão da palavra em letras ou sílabas, divisão silábica), a análise da palavra em contexto (categoria sintáctica ou informação semântica) ou derivadas da análise da relação dessa palavra com as restantes existentes no léxico (p. ex. frequência de uso da palavra na escrita e/ou na fala, similaridade ortográfica ou fonológica com outras palavras, frequência de bigrama, etc.), como propriedades de natureza mais subjectiva que implicam a recolha de medidas que reflectem as experiências pessoais dos indivíduos com o uso da própria língua (p. ex. idade-de-aquisição, imaginabilidade, familiaridade, concreteza, emocionalidade).

A manipulação sistemática destes atributos na investigação tem contribuído de forma decisiva não só para a compreensão da arquitectura e processamento linguístico humano, como para a compreensão do funcionamento de outros sistemas cognitivos como a memória, a atenção, a representação mental de conceitos ou a compreensão de determinados processos desenvolvimentais (p. ex. aquisição da fala, leitura) tanto em populações "normais" como em populações com trajectórias atípicas de desenvolvimento. Refira-se também que os contributos associados a este

tipo de ferramentas não se limitam ao seu uso como instrumento de apoio à investigação, mas também como um meio para obter um conhecimento mais aprofundado das características da própria língua. Com efeito, não apenas a criação mas também a disponibilização pública deste tipo de recursos é importante e urgente, especialmente quando se compara com os recursos existentes para outras línguas. Assim, na linguística descritiva, será uma ferramenta útil para a análise e descrição fonológica, morfossintáctica e semântica do PE, particularmente na análise quantitativa. Poderá vir a ser também um recurso muito importante para a Linguística aplicada (por exemplo, para a Lexicografia e a Terminologia do PE), fornecendo informação sobre o uso real de palavras e acepções, bem como a sua frequência, etc., assim como para a análise estilística (não apenas do ponto de vista literário, mas também pedagógico, forense, sóciolinguístico, cultural, etc.), nomeadamente graças ao trabalho de etiquetação realizado (com índices objectivos e subjectivos). Em suma, permitirá realizar estudos com base em informação descritiva, estatística e classificativa que anteriormente não estava disponível, designadamente numa única plataforma.

Para o PLN esta base de dados poderá ser utilizada em diversas vertentes, desde a simples correcção ortográfica (tendo em conta vizinhança ortográfica e fonética, por exemplo), à síntese de voz (dada a inclusão de transcrição fonética) e à análise semântica, dado o interesse do P-Pal em integrar relações semânticas.

Em Portugal, o reconhecimento da necessidade deste tipo de bases é relativamente recente. Assim, e embora tais bases se encontrem disponíveis em línguas como o inglês (p. ex. MRC (Coltheart, 1981); N-Watch (Davis, 2005); E-Lexicon (Balota et al., 2007)), o francês (p ex. BRULEX (Content, Mousty e Radeau, 1990); LEXIQUE (New et al., 2001; New et al., 2004); French Lexicon Project(Ferrand et al., 2010)), o holandês e o alemão (p. ex. CELEX (Baayen, Piepenbrock e Gulikers, 1995; Baayen, Piepenbrock e van Rijn, 1993)), o grego (p. ex. GreekLex (Ktori, van Heuven e PitchHford, 2008)), ou o espanhol (p. ex. LEXESP (Sebastián-Gallés et al., 2000); BuscaPalabras (Davis e Perea, 2005)), elas são praticamente inexistentes para o português. Até aos anos 90, o indicador psicolinguístico mais citado pelos investigadores nacionais era o de frequência de uso das palavras num trabalho designado Português Fundamental (Nascimento, Marques e da Cruz, 1987) e baseado num corpus oral de pequenas dimensões (700.000 palavras). Embora nos últimos anos se tenha reconhecido essa limitação e se tenham desenvolvido esforços no sentido de construir bases lexicais que contivessem outros indicadores linguísticos importantes, a verdade é que elas apresentam um número muito reduzido de informações. Para além da informação ortográfica disponível em todas elas (e que configura as suas entradas lexicais), cada uma contém apenas informação relativa ou à transcrição fonética ou à caracterização morfossintáctica das palavras (Nascimento, Rodrigues e Gonçalves, 1996).

Procurando ultrapassar tais dificuldades surgiu a PORLEX (Gomes e Castro, 2003). A PORLEX é uma base lexical que reúne informações de tipo ortográfico, fonológico, fonético, gramatical e de vizinhança para um total de 29.238 palavras e que constitui um instrumento útil à investigação cognitiva em geral e à da psicolinguística em particular. Contudo, as limitações que apresenta ao nível do valor de frequência lexical que disponibiliza (importado do trabalho Português Fundamental que, para além de se revelar desactualizado, apenas é disponibilizado para cerca de 5% das suas entradas lexicais) impedem um uso mais alargado dessa ferramenta na investigação nacional. Ora, na actualidade, o PE conta já com novos léxicos de frequências extraídos de corpora de grandes dimensões (p. ex. CORLEX (Nascimento, Pereira e Saramago, 2000)) e de vários corpora como o CETEMPúblico, o ECI-EE, o FrasesPP, os Clássicos da Porto Editora, o Natura/Minho, o Vercial, o Avante e o Dia-CLAV disponíveis na rede, no sítio da Linguateca[1] (Costa, Santos e Cardoso, 2008). Não obstante, embora disponibilizem informação de frequência de uso mais actualizada, diversificada e representativa, não disponibilizam outras informações sobre outras propriedades lexicais das palavras, como a PORLEX. Urge assim desenvolver novas aplicações que incorporem todas estas informações numa única ferramenta.

No que se refere aos índices psicolinguísticos subjectivos, alguns autores, reconhecendo também essa lacuna nas bases nacionais e a sua relevância na investigação cognitiva e neurocognitiva mais actual, desenvolveram estudos que procuraram avaliar a familiaridade (Garcia-Marques, 2003; Marques, 2004), a valência (Garcia-Marques, 2003), a imaginabilidade e a concreteza (Marques, 2005), e a idade de aquisição (Cameirão e Vicente, 2010; Marques et al., 2007) de palavras portuguesas. Contudo, apesar da relevância desses trabalhos a verdade é que eles incidiram sobre um número bastante

[1] `http://www.linguateca.pt/ACDC/`

restrito de palavras (p. ex. 459 para o índice de familiaridade e 249 para o índice de imaginabilidade (Marques, 2004; Marques, 2005)) e, mesmo para aqueles que avaliaram os mesmos índices, a adopção de procedimentos de avaliação distintos (veja-se, por exemplo, a forma como a variável familiaridade é avaliada nos estudos de Garcia-Marques (2003) e Marques (2004); ou a idade de aquisição nos estudos Cameirão e Vicente (2010) e Marques (2005)) impede a sua utilização conjunta.

Por último, o suporte informático em que se apresentam (Microsoft Excel), embora garanta alguma flexibilidade de pesquisa, a verdade é que pode dificultar a selecção de estímulos quando, como na maioria das vezes acontece, o investigador pretende controlar um conjunto diversificado de parâmetros relativos às palavras ao mesmo tempo. Além disso, dado que as informações das palavras se encontram em suportes distintos, o investigador terá sempre de recorrer a distintas aplicações informáticas para seleccionar os estímulos apropriados, correndo sempre o risco de, nas diferentes aplicações, não encontrar as mesmas entradas lexicais. Assim, e independentemente do paradigma experimental adoptado ou da área de investigação considerada, os investigadores portugueses deparam-se na actualidade com sérias dificuldades no planeamento e condução dos estudos que utilizem estímulos verbais, e, em geral, na análise e descrição linguística do PE baseadas em corpora. Com o presente projecto pretendemos colmatar essa necessidade desenvolvendo uma aplicação informática multi-plataforma designada Procura-PALavras (P-PAL) que, com comodidade e rapidez, permita calcular, em simultâneo, um conjunto de índices psicolinguísticos objectivos e subjectivos para palavras do PE, num formato amigável e disponibilizado gratuitamente à comunidade científica a partir de um sítio em linha a construir para o efeito.

2 Procura-PALavras (P-PAL)

O P-PAL será a versão adaptada para o PE do *software* inglês N-Watch (Davis, 2005) já adaptado para o espanhol como BuscaPalabras (Davis e Perea, 2005) e Basco como E-Hitz (Perea et al., 2006) considerando as características particulares do sistema do PE contemporâneo. Permitirá, para além da computação do valor de frequência por milhão e logarítmico (base 10) de todos os lemas e formas que constituirão as suas entradas lexicais (indexadas a partir da compilação, tratamento e análise de vários corpora recentes), a realização de um conjunto diversificado de análises relativas quer às dimensões morfológicas e morfo-sintácticas (p. ex. classe gramatical, número de morfemas, frequências por tipo, ocorrência, forma e lemas por classe, género e número); quer às dimensões ortográficas (p. ex. número de letras, estrutura consoante-vogal, ponto de unicidade, homógrafos e diversas medidas de frequências por tipo e ocorrência de bi e trigramas e de vizinhanças); fonológicas (p. ex. pronúncia da palavra, número de fonemas, vogais neutras, homófonos e diversas medidas de frequências tipo e ocorrência de bifone e de vizinhanças); silábicas (p. ex. silabificação ortográfica e fonológica da palavra, número de sílabas, estrutura silábica, padrão de acento e diversas medidas de frequências tipo e ocorrência de vizinhanças silábicas ortográfica e fonológica); e semânticas (p. ex. número de acepções da palavra, co-ocorrências e distância semântica) de palavras do PE. Permitirá ainda obter índices para pseudo-palavras (que, a par das palavras, constituem estímulos de ampla utilização nos diferentes paradigmas da investigação experimental), e para os índices subjectivos de imaginabilidade, concreteza, familiaridade, valência, activação e controlabilidade, ainda não disponíveis entre nós ou, como vimos, disponíveis para um léxico bastante restrito.

Ao utilizar o P-PAL o utilizador poderá assim fazer uma utilização personalizada do programa ao seleccionar, da ampla variedade de análises disponíveis aquelas que se adequam aos propósitos da sua investigação e numa dupla possibilidade de utilização: o utilizador poderá optar por pedir ao programa que avalie um conjunto de palavras previamente definidas pelo investigador num conjunto de parâmetros seleccionados do menu de análises (p. ex. frequência lexical, número de letras, estrutura consoante-vogal, vizinhos ortográficos por substituição, adição e subtracção, frequência das formas dos vizinhos de frequência alta, distância de Levenshtein) ou poderá pedir ao programa que lhe faculte as palavras que, entre as que fazem parte da base lexical, obedeçam a esses parâmetros. Cremos que esta característica da ferramenta, não disponível na versão original do N-Watch (Davis, 2005), do BuscaPalabras (Davis e Perea, 2005) ou do E-Hitz (Perea et al., 2006) oferece maior versatilidade à ferramenta. O P-PAL assume-se assim como uma ferramenta de investigação fundamental e indispensável à promoção e internacionalização da investigação em Portugal.

3 Fases de execução do projecto

O projecto P-PAL é um projecto claramente interdisciplinar onde os contributos das áreas da Psicolinguística, da Linguística e do Processamento de Linguagem Natural (PLN) se assumem como essenciais à sua execução. Embora tais contributos sejam importantes ao longo de todo o projecto, podemos distinguir três fases principais que configuram o contributo mais acentuado de alguma delas em cada momento temporal da sua implementação.

Assim, a primeira fase do projecto, já em curso (a decorrer entre Maio de 2010 e Maio 2011), envolverá essencialmente o contributo da área da Linguística e do PLN na constituição do vocabulário por defeito a incluir no P-PAL (e que consubstanciarão as suas entradas lexicais - lemas e formas) e na extracção dos seus valores de frequência lexical (absoluta, por milhão e logarítmica – base 10). Tal tarefa compreenderá a recolha, o tratamento e a análise de vários corpora recentes do PE de diversos géneros literários e dimensões com informação de frequência de uso disponível. Ainda durante este primeiro ano de execução do projecto levar-se-á a cabo a inserção semi-automática da informação linguística estrutural das entradas lexicais do P-PAL (p. ex. informação morfo-sintáctica, transcrição fonética, silabificação, padrão de acento), a verificação e correcção da base, e a selecção do conjunto de palavras sobre as quais se recolherão medidas subjectivas. Dar-se-á também inicio à construção do interface e da aplicação na rede a partir dos quais se disponibilizarão os índices à comunidade de investigadores.

A segunda fase do projecto (Maio de 2011 – Maio 2012), envolverá essencialmente o contributo das áreas do PLN na computação das métricas de frequências por tipo e ocorrência e de vizinhanças de cada um dos índices integrados no P-PAL (índices ortográficos, fonológicos, fonográficos, silábicos ortográficos e fonológicos), e da Psicolinguística na preparação dos materiais e procedimentos na recolha presencial, lápis-papel, e a recolha via aplicação na rede, dos índices subjectivos a incluir na base (familiaridade, imaginabilidade, concreteza, valência, activação e controlo).

A terceira e última fase do projecto (Maio de 2012 – Maio 2013), envolverá essencialmente o contributo das áreas do PLN na computação das métricas semânticas a incluir na base e na computação de métricas para pseudo-palavras (frequências por tipo e ocorrência de bigrama e trigrama e de vizinhanças ortográficas e fonológicas), e da Psicolinguística na conclusão da recolha e no tratamento dos índices subjectivos a incluir no P-PAL.

4 Conclusão

O Procura-PALvras (P-PAL) é um projecto interdisciplinar que cruza as áreas da Psicolinguística, da Linguística e do Processamento de Linguagem Natural (PLN) na construção de uma ferramenta electrónica que habilite os investigadores nacionais com um instrumento que funcione ora como um meio de apoio à investigação nas diferentes áreas do questionamento científico (p. ex. Psicologia Cognitiva, Neurociências, Linguística, PLN), ora como um meio para um conhecimento mais aprofundado das características da própria língua e para o apoio ao desenvolvimento de aplicações capazes de processar a linguagem natural.

Pela inovação que constitui entre nós, pela diversidade de índices que aglutina (índices de frequência lexical, índices morfológicos e morfo-sintácticos, índices ortográficos, índices fonológicos, índices fonográficos, índices silábicos ortográficos e fonológicos, índices semânticos, índices subjectivos e índices para pseudo-palavras) e pela dupla funcionalidade de análises que oferece ao utilizador (avaliar palavras em determinados parâmetros e obter palavras que obedeçam a tais parametros), consideramos estar perante uma ferramenta com um potencial inestimável à promoção e 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 programa COMPETE (Programa Operacional Factores de Competitividade), integrado no Fundo Europeu de Desenvolvimento Regional (FEDER), o financiamento deste projecto (PTDC/PSI-PCO/104679/2008).

Referências

Baayen, Harald R., Richard Piepenbrock, e Leon Gulikers. 1995. *The CELEX Lexical Database. Release 2 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Baayen, Harald R., Richard Piepenbrock, e H. van Rijn. 1993. *The CELEX Lexical Database. Release 1 (CD-ROM).* Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania.

Balota, David A., Melvin J. Yap, Michael J. Cortese, Keith I. Hutchison, Brett Kessler,

Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, e Rebecca Treiman1. 2007. The english lexicon project. *Behavior Research Methods*, 39:445–459. http://artsci.wustl.edu/~rtreiman/Selected_Papers/English_Lexicon_Project_userguide_in%20press.pdf.

Cameirão, Manuela L e Selene G. Vicente. 2010. Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior Research Methods*, 42(2):474–480.

Coltheart, Max. 1981. The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33A:497–505.

Content, Alain, Phillppe Mousty, e Monique Radeau. 1990. Brulex: une base de données lexicales informatisée pour le français écrit et parlé. *L'année psychologique*, 90:551–566. http://www.lexique.org/public/Brulex.pdf.

Costa, Luís, Diana Santos, e Nuno Cardoso. 2008. Perspectivas sobre a Linguateca / Actas do encontro Linguateca : 10 anos, 11 de Setembro, 2008. http://www.linguateca.pt/LivroL10/Livro-Costaetal2008.pdf.

Davis, Colin J. 2005. N-Watch: a program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, 37(1):65–70. http://www.pc.rhul.ac.uk/staff/c.davis/Articles/Davis_05.pdf.

Davis, Colin J. e Manuel Perea. 2005. BuscaPalabras: a program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in spanish. *Behavior Research Methods*, 37(4):665–671. http://brm.psychonomic-journals.org/content/37/4/665.full.pdf.

Ferrand, Ludovic, Boris New, Marc Brysbaert, Emmanuel Keuleers, Patrick Bonin, Alain Méot, Maria Augustinova, e Chistophe Pallier. 2010. The French lexicon project: Lexical decision data for 38,840 French words and 38,840 pseudowords. *Behavior Research Methods*, 42(2):488–496. http://www.mariaaugustinova.com/site/publications_files/FERRAND-BRM-Final-2010.pdf.

Garcia-Marques, Teresa. 2003. Avaliação da familiaridade e valência de palavras concretas e abstractas em língua portuguesa. *Laboratório de Psicologia*, 1(1):21–44. http://repositorio.ispa.pt/bitstream/10400.12/124/1/LP%20%281%291%20-%2021-44.pdf.

Gomes, Inês e São Luís Castro. 2003. Porlex: A lexical database in European Portuguese. *Psychologica*, 32:31–108. http://www.fpce.up.pt/labfala/porlex_gomes&castro03.pdf.

Ktori, Maria, Walter J. B. van Heuven, e Nicola J. PitcHford. 2008. GreekLex: A lexical database of modern Greek. *Behavior Research Methods*, 40(3):773–783. http://brm.psychonomic-journals.org/content/40/3/773.full.pdf+html.

Marques, J. Frederico. 2004. Normas de familiaridade para substantivos comuns. *Laboratório de Psicologia*, 2:5–19.

Marques, J. Frederico. 2005. Normas de imagética e concreteza para substantivos comuns. *Laboratório de Psicologia*, 3:65–75.

Marques, J. Frederico, Francisca L. Fonseca, A. Sofia Morais, e Inês A. Pinto. 2007. Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables. *Behavior Research Methods*, 39(3):439–444. http://brm.psychonomic-journals.org/content/39/3/439.full.pdf.

Nascimento, Maria Fernanda Bacelar, M. Lúcia Garcia Marques, e M. Luísa Segura da Cruz. 1987. *Português Fundamental: Métodos e documentos (Vol. II, Tomo I: Inquérito de frequência)*. INIC, Centro de Linguística da Universidade de Lisboa, Lisboa.

Nascimento, Maria Fernanda Bacelar, Luísa Pereira, e João Saramago. 2000. Portuguese corpora at CLUL. Em *Second International Conference on Language Resources and Evaluation*, volume II, pp. 1603–1607, Athens.

Nascimento, Maria Fernanda Bacelar, Maria Celeste Rodrigues, e José Bettencourt Gonçalves, editores. 1996. *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística*, volume I: Corpora, Lisboa. Colibri.

New, Boris, Christophe Pallier, Marc Brysbaert, Ludovic Ferr, Royal Holloway, U Service, e Hospitalier Frédéric Joliot. 2004. Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, 36:516–524.

New, Boris, Christphe Pallier, Ludovic Ferrand, e Rafael Matos. 2001. Une base de

données lexicales du Français contemporain sur internet: LEXIQUE. *L'Année Psychologique*, 101:447–462. `http://www.pallier.org/papers/Lexique.2001.pdf`.

Perea, Manuel, Miriam Urkia, Colin J. Davis, A. Agirre, E. Laseka, e M. Carreiras. 2006. E-Hitz: A word-frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, 38:610–615. `http://www.uv.es/~mperea/ehitz.pdf`.

Sebastián-Gallés, Núria, Maria Antònia Martí Antonín, Manuel Francisco Carreira Valinã, e Fernando Cuetos Vega. 2000. *LEXESP: Léxico informatizado del español*. Edicions de la Universitat de Barcelona, Barcelona.