Referência:
Gomes, I., & Castro, S. L. (2003). Porlex, a lexical database in European Portuguese. *Psychologica, 32*, 91-108.

Porlex, a lexical database in European Portuguese

Inês Gomes[1] and São Luís Castro[2]

[1]Professor Auxiliar, Universidade Fernando Pessoa, Porto, Portugal
[2]Professor Associado, Faculdade de Psicologia e Ciências da Educação, Universidade do Porto, Porto, Portugal

Address for correspondence:
São Luís Castro
FPCE-Universidade do Porto
Rua do Campo Alegre, 1021
P 4169 - 004 Porto Portugal
Phone: +351 22 607 9756; Fax: +351 22 607 9725; E-mail: slcastro@psi.up.pt

Running head: PORLEX DATABASE IN EUROPEAN PORTUGUESE

Abstract

This paper presents a tool for research in the psychology of language, a computerized lexical database in European Portuguese. Porlex was built on the basis of a middle sized adult lexicon, and provides orthographic, phonological, phonetic, part-of-speech, and neighborhood information for about 30 000 words (uninflected content words and inflected function words). Frequency was included whenever possible. After highlighting the role of lexical databases for experimental research on language, we give an overview of the sources and contents of Porlex 1.0, with a special emphasis on the 44 different types of information it provides. A brief characterization of the corpus is also included.

Key words: Database, Lexicon, European Portuguese, Psychology of Language

Lexical Databases as Research Tools

Contemporary research about language can hardly be done without an elaborate and sometimes laborious process of selection of stimuli. Words, syllables or phonemes that are presented to the participants have to be carefully chosen not only on the basis of the variables under scrutiny, but also depending on a number of other characteristics that may affect the perceptual and cognitive processes involved in the preparation of the response. Words that differ on semantic category may have to be matched on frequency, imageability, or syllabic structure; syllables that differ on structure may have to be chosen according to their relative frequency in the language or potential position in the word; phonemes that differ on word position may have to be matched on orthographic consistency. In order to accomplish this, an impressive amount of knowledge is required, namely knowledge on aspects of language that are, or are presumed to be, cognitively relevant. A further source of complication is that language as such is a highly abstract entity. In practice, what the researcher deals with are specific languages, such as English, Portuguese, French, etc. Thus, the need of a reliable source of information that is both language specific and cognitively founded has become apparent with the advances of experimental and neuropsychological language research. Computerized lexical databases are a response to this need.

Over the last years, several databases have been developed. One of the first was the MRC Psycholinguistic Database (Coltheart, 1981), that gathered psycholinguistic measures on 150 000 English words. Other languages followed; for example, French, with Brulex (Content, Mousty, & Radeau, 1990) and Lexique (New, Pallier, Ferrand, & Matos, 2001), Dutch and German with Celex (Baayen, Piepenbrock, & Gulikers, 1995), and Spanish (Piñeiro & Manzano, 2000). Most were developed from adult vocabularies, but others are based on children's lexica (e.g., Piñeiro & Manzano, ib., Lambert & Chesnet, 2001).

Typically, these databases provide orthographic, phonological, grammatical and frequency information on either lemmas (uninflected wordforms) or inflected wordforms, but they vary widely on format and size. For example, Brulex has ca. 30 000 lexical entries, most of them lemmas (verbs in the infinitive; content words in the singular form), and some inflected words (articles and pronouns; non-homophonic plural and feminine forms, e.g., *journal/journaux; petit/petite*); 29 informations are given for each entry, ranging from phonological transcription to neighborhood count. Lexique is composed of three related files, one based on written inflected wordforms (ca 130 000 entries), another on lemmas (ca 55 000 entries), and the third compiling frequency measures of letters, bigrams, trigrams, phonemes and syllables of the inflected wordforms. Celex is a trilingual database, that comprises three separate lexica in Dutch, English and German: lemmas, wordforms, and the corpus lexicon that combines both. The number of entries depends on type of lexicon and language, from ca 52,000 German lemmas to 380,000 Dutch wordforms, with English in-between – ca 53,000 lemmas and 160,000 wordforms. Approximately 950 different types of linguistic and psycholinguistic informations are provided for each entry (Burnage, 1990; Piepenbrock, 2001).

In a recent review, Nascimento, Rodrigues and Gonçalves (1996) listed 26 corpora in European Portuguese (cf. also http://www.clul.ul.pt/). About half are lexical databases that contain written words; some include morpho-syntactic annotations, others phonetic transcriptions. However, none provides the range of informations that are required for cognitively-oriented research on language (cf. Appendix A). This situation lead us to develop Porlex.

Porlex: Sources, Overview, Corpus and Variables

Porlex is a computerized lexical database in European Portuguese designed as a tool for research in the psychology of language. It was built on the basis of a middle sized adult lexicon, and provides orthographic, phonological, part-of-speech, and neighborhood information for each entry. Its current version, Porlex 1.0, contains 29,238 different words and 44 types of information. Porlex 1.0 is in Excel format (Microsoft Corporation, 1998). It is available for non-commercial purposes upon request to the authors.

Sources for Porlex

Lexical entries, grammatical and morphological classification, syllabication, phonetic transcription and frequency information were collected from various sources. There are listed on a separate section of the References, and will be briefly reviewed here. Porlex words come from the Dicionário Universal Fundamental (Texto Editora, 1998), that was selected because of its size. In the process of compiling the remaining source informations, it became clear that the final selection of the lexical entries had to be fine tuned. For that purpose, we used Porto Editora (Costa & Melo, 1997) and Cândido de Figueiredo dictionaries (1996), as well as the grammars of Cunha and Cintra (1987), Mateus, Brito, Duarte and Faria (1989), and Vilela (1995). The grammars were specifically used to enter inflected pronouns and contractions, as well as prepositions and conjunctions. Morphological classification and grammatical class were also compiled from these sources. Syllabication of the orthographic wordforms was marked according to Michaelis dictionary (Melhoramentos, 1998). Phonetic transcriptions were based on the pocket Langenscheidts Portuguese dictionary (Irmen & Kollert, 1995) and on Vilela (1991). Frequency information comes from Nascimento, Marques and Cruz (1987) and Nascimento, Rivenc and Cruz (1987), the only source of frequency information available to us during the period while Porlex was being compiled. The "Dicionário da Língua Portuguesa Contemporânea" (Academia das Ciências de Lisboa, 2001), that includes phonetic transcription, and CORLEX, a more recent source of frequency information (Nascimento, Casteleiro, Marques, Barreto, & Amaro, n.d.) were available shortly afterwards[1].

Porlex entries were inserted automatically whenever possible. However, because we were unable to find source information in compatible electronic format, about a third of these was typed in manually. For example, at the time Porlex was started there were no Portuguese middle-sized dictionaries that included phonetic transcription of the words, and these had to be entered individually. A comprehensive survey of Porlex contents follows.

Overview

From the source informations, we extracted automatically almost all of the remaining Porlex contents (cf. Table 1). Some computations, like the number of letters in a word, were performed on a single entry. Others were structural, in that they explored the relations between words; these were computed on multiple entries.

Insert Table 1 about here

The orthographic representation is a good starting point for a lexical database, because in all likelihood it is the most robust way to represent a word. Being based on widely accepted conventions that are hardly ever updated, the written form of the word rarely changes. Accordingly, Porlex is based on the orthographic wordforms, that dictate the alphabetical order by which words are organized. Several informations are specifically related to this wordform: whether it contains diacritics, how it is divided into syllables, and how many letters and syllables it contains (Variables 3, 10, 11 and 16, respectively), among others.

The representation of the spoken word, however, is more prone to variability. In Porlex, we present two transcriptions for each word, the phonetic wordform that differentiates between allophones (e.g., different symbols for the first and last segment of *lua* and *sal*, respectively), and the phonemic wordform where that level of phonetic detail is absent (see next section for a fuller explanation). The symbols used in the transcriptions are presented in Appendix B. Due to the lack of a well established source for the transcription of spoken words in Standard European Portuguese, and also because it is a useful information as such, Porlex includes a variable that signals the words for which more than one broad phonetic transcription is acceptable (cf. infra Variable 6, Variant Pointer), and the alternative phonetic transcriptions are given as separate entries (for a maximum of three different alternatives, cf. Variables 40 to 42). Porlex gives additional informations about the characteristics of spoken words; for example, how many schwas they include, how they are divided into syllables, where is the stressed syllable; also a pointer for ambisyllabicity, the length in number of phonemes or of syllables, and two different types of phonetic patterns, one based on a gross classification of speech sounds, another more distinctive. For some of these, the difference between phonetic and phonemic wordforms is irrelevant. In such cases, we use the label phonological (cf. Variables 14 and 17).

Irrespective of whether their form is written or spoken, words fall into different type-of-speech classes. Porlex provides this information in Variables 7 and 8, where words are classified according to grammatical class, and as open or closed, respectively. The grammatical gender of the word is also given; because nouns and adjectives are uninflected, we added a variable that signals whether gender inflexion is permissible and another that marks plural forms (Variables 19 to 21).

Porlex also provides structural measures on the similarities between words, that were extracted by comparing the characteristics of a given lexical entry with the remaining Porlex entries. One set of such measures deals with neighborhood similarities: neighborhood densities, uniqueness points and the listing of the neighbors of a given target were computed for orthographic, as well as for phonetic and for phonemic wordforms (Variables 23 to 31). Neighbors were defined according to Luce (1986; also Charles-Luce & Luce, 1990), as words that differ on a single phoneme either by substitution or by deletion/addition when relative serial position of the segments is maintained, and uniqueness points according to Marslen-Wilson and Tyler (1980; also Marslen-Wilson, 1990) as the position of the segment that discriminates a word from its neighbors, counted from the beginning. The second set of structural computations involved the comparison between orthographic and phonetic wordforms in order to extract the number of nonhomophonic homographs, homographic homophones and nonhomographic homophones, that are presented in Variables 32 to 34. For a detailed presentation of the computational procedures, see Gomes (2001).

Corpus

As a tool for research in the psychology of language, Porlex should incorporate words presumably represented in the mental lexicon. This thought guided the criteria that were adopted to fine-tune the selection of lexical entries into Porlex. For practical reasons we had to start with lemmas rather than wordforms, but in the case of articles and pronouns should feminine or plural forms be left out, like, say, the word 'we'? Based on the well established distinction between content vs. functional words (e.g., Fromkin & Rodman, 1998; Segalowitz & Lane, 2000), we decided to include the inflected forms of functional words, and as many different forms of this type of words as possible. Grammars were chosen as the primary source for the inclusion of all forms of articles, pronouns, prepositions, contractions and conjunctions into Porlex. For the remaining word types, dictionaries were used as the source for Porlex entries.

Insert Table 2 about here

The main characteristics of the resulting corpus are summarized in Table 2. Nouns account for 56% of the words; adjectives and verbs are the second largest categories, each with ca 20% of the corpus. Even with the criterion of including all forms (inflectionally and lexically) of articles, pronouns, contractions, prepositions and conjunctions, and only the noninflected forms of nouns, adjectives and verbs, the difference in the number of entries for these categories is huge. Adverbs account for 4% of Porlex. Almost 90% of them are derived from adjectives by addition of the suffix *–mente* (equivalent to *–ly*). For this reason (cf. Azuaga, 1996), they were classified as open class words, together with nouns, adjectives and verbs. These account for 98% of the corpus. There is an almost even distribution of nouns into feminine vs. masculine forms, and ca 2% of them are in the plural form.

Variables

In this section, we present a synopsis of the different variables in Porlex. For each, we show Number, Code Name and Name proper, Values (examples for string type variables; codes for classification variables; and range for numeric variables), and a brief explanatory definition (Concept).

1. # - Number

VALUES**:** *1, 2, . . ., 29 238.*

CONCEPT**:** Entry number according to alphabetical order of the orthographic wordforms. Homographs are ordered by grammatical class (see variable 7, Grammatical Class) whenever possible.

2. Orto - Orthographic Wordform

VALUES: *afável, alpercata, boneca, crítico, lenha, musgo, sala*, . . .

CONCEPT: Orthographic wordforms in lower case including diacritics. These are the primary entries of Porlex. They consist of uninflected content words and inflected function words of a middle sized adult lexicon. By convention, only two instances of orthographically identical entries were accepted: nonhomophonic homographs, and homophonic homographs differing in grammatical class. E.g, *bola* pronounced with an open [o] and *bola* with a closed [o] (ball vs. round yeast bread, respectively); the homophonic homographs *crítico,* noun, and *crítico*, adjective (critic vs. critical).

3. Diacr - Diacritic Pointer

VALUES: *1* = at least one diacritic; *0* = no diacritics.

CONCEPT: Code for the presence of diacritics in the orthographical wordform.

4. Fot - Phonetic Wordform

VALUES: *A´favE9, a9p6r'katA, bu´nEkA, ´kritiku, ´lANA, ´muZgu,´salA*, . . .

CONCEPT: Phonetic transcription of the orthographic wordform using an adapted Unibet system, by reference to a careful pronunciation of the corresponding isolated word. The following criteria were adhered to: schwa always transcribed, as well as allophonic variation relative to /l/ velarization, assimilation of /S/, fricatization of voiced stops, and coarticulatory homorganic nasals; as appropriate, either diphthongation or crasis of orthographic vowel sequences.

5. Fom - Phonemic Wordform

VALUES: *A´favEl, alp6r'katA, bu´nEkA, ´kritiku, ´lANA, ´muSgu, ´salA*, . . .

CONCEPT: Broader phonetic transcription using the same Unibet system; the schwa is maintained, as is diphthongation or crasis of vowel sequences, but allophonic variation is removed (velarization of /l/, assimilation of /S/, fricatization of voiced stops, and coarticulatory homorganic nasals not marked).

6. Var - Variant Pointer

VALUES: *1* = phonetic variant *2* = lexical variant; *3* = both; [empty] if no variants.

CONCEPT: Code for whether the orthographic wordform has more than one acceptable pronunciation, or can be written in a slightly different way (phonetic and lexical variant, respectively), or both. Phonetic variants include minor dialectal variations (also acceptable as standard pronunciations). Lexical variants are orthographically similar synonymous words, e.g., *raquete* vs. *raqueta*. They were not included as separate entries because they might increase artefactually the number of neighbors of a given word.

7. CGram - Grammatical Class

VALUES: *no* = noun; *aj* = adjective; *vb* = verb; *av* = adverb; *nu* = numeral; *pr* = preposition; *ar* = article; *pn* = pronoun; *ct* = contraction; *co* = conjunction; *in* = interjection; *lo* = locution.

CONCEPT: Code for grammatical word class, according to a classification into: noun, adjective, verb, adverb, numeral, preposition, article, pronoun, contraction, conjunction, interjection and locution. Whenever applicable, this order is used because it facilitates the extraction of same class words and the categorization into open vs. closed class words (see below).

8. CAF - Open/Closed Class

VALUES: *a* = open class; *f* = closed class; *nu* = numeral; *in* = interjection; *lo* = locution.

CONCEPT: Code for the classification of word types into open class vs. closed class or functional word. Nouns, adjectives, verbs, and adverbs were classified as open class words. Prepositions, articles, pronouns, conjunctions, and contractions were classified as functional words. The remaining word types retain their classification from the Grammatical Class variable (as numerals, interjections, or locutions).

9. Tonic - Stress Position

VALUES: *ag* = oxytone; *gr* = paroxytone; *es* = proparoxytone; [empty] if monosyllabic word.

CONCEPT: Code for the position of the stressed syllable in multisyllabic words: in the last, penultimate, or antepenultimate syllable (oxytone, paroxytone and proparoxytone words, respectively). Position established by reference to the variable 17, Phonological Syllabication.

10. Nlet - Letter Length

VALUES: *1, 2, . . ., 22*.

CONCEPT: Number of letters in the orthographical wordform.

11. NSilO - Orthographic Syllable Length

VALUES: *1, 2, . . ., 10*.

CONCEPT: Number of syllables in the orthographic wordform, computed by reference to variable 16, Orthographic Syllabication (see below).

12. Nfom - Phonemic Length

VALUES: *1, 2, . . ., 20*.

CONCEPT: Number of segments in the phonemic wordform. This is equivalent to the number of phones when the phonetic wordform does not include homorganic nasals. Phonetic length (including homorganic nasals) can be computed in conjunction with the following variable.

13. NHC - N Homorganic Nasals

VALUES: *1, 2, 3, 4*; [empty] if no homorganic nasals.

CONCEPT: Number of coarticulatory homorganic nasals in the phonetic wordform.

14. NSilF - Phonological Syllable Length

VALUES: *1, 2, . . ., 10*.

CONCEPT: Number of syllables in the spoken wordform, computed by reference to the variable 17, Phonological Syllabication. The distinction between phonetic vs. phonemic wordforms is irrelevant for syllable segmentation and syllable count.

15. NSchwa - N Schwa

VALUES: *1, 2, . . ., 5*; [empty] if no schwas.

CONCEPT: Number of schwas in the spoken wordforms (applies to phonetic as well as to phonemic transcriptions).

16. DivSilO - Orthographic Syllabication

VALUES: *a-fá-vel, al-per-ca-ta, bo-ne-ca, crí-ti-co, le-nha, mus-go, sa-la, . . .*

CONCEPT: Each non-final syllable of the orthographic wordform is separated from the following syllable by an hyphen. Monosyllables are not marked. Syllables boundaries are marked according to translineation rules. Within-syllable letter sequences are the following: CV, V and the digraphs *ch, lh, nh, gu, qu* (e.g., *cha-ve*, *pa-lho-ta*, *san-gues-su-ga*, *fran-qui-a*); vowel strings corresponding to dipthongs or tripthongs (*ou-ro*, *sa-guão*); CC in initial word position, Cr, and Cl in monomorphemic words (*pneu-má-ti-co*, *su-bli-me*; but *sub-lo-ca-ção*); (C)Vm and (C)Vn if the vowel is nasal (*an-dar, am-bâr*); (Cl,r)Vr, (Cl,r)Vl, (Cl,r)Vs and (Cl,r)Vz (e.g., *al-gar, mar-mo-ta*, *a-ves-truz*); and (C)VCs followed by a consonant (*abs-tra-ir*). Between-syllable letter sequences are: adjacent vowels forming a hiatus, adjacent consonants when one of them is part of an affix, and consonant sequences not described above, including the digraphs *ss* and *rr* that are separated by convention (*sa-í-da*, *rap-to*, *ob-jec-ti-vo*, *pás-sa-ro*).

17. DivSilF - Phonological Syllabication

VALUES: *A΄fa.vE9,a9.p6r'ka.tA, bu΄nE.kA, ΄kri.ti.ku, ΄lA.NA, ΄muZ.gu, ΄sa.lA, . . .*

CONCEPT: Each non-final, nonstressed syllable of the multisyllabic phonetic wordforms is separated from the preceding syllable by a dot. The stressed syllable is marked according to the usual convention. Syllable boundaries are defined according to criteria analogous to the ones described for orthographic forms, after matching digraphs into single phonemes (e.g., *carro* = ['ka.Ru]), and assimilating homorganic nasals to the preceding nasal vowels. Thus, syllable boundaries are equivalent in the phonetic and in the phonemic wordforms. Potential ambisyllabicity is indicated by a pointer in a separate variable (see below). The following criteria are used to divide ambisyllabic strings: in VGV sequences, the glide is grouped with the second vowel into a separate syllable; in CC sequences the two consonants are separated into adjacent syllables (*['pra.jA]*, *[Op'tar]*).

18. AmbSil - Ambisyllabicity Pointer

VALUES: *a* = alternative segmentation into phonetic syllables is possible; [empty] if not.

CONCEPT: Code for ambisyllabicity, that may occur at least in Vowel-Glide-Vowel sequences, as in the word *maio*, *['maju]*, and in Consonant-Consonant sequences when the second consonant is different from [r] or [l], as in *septo*, *['sEptu].*

19. Gen - Gender

VALUES: *f* = feminine; *m* = masculine; *h* = invariant; [empty] if gender not applicable.

CONCEPT: Code for the classification of nouns, adjectives, numerals and pronouns according to grammatical gender: feminine, masculine, or invariant. The remaining word types are not classified.

20. FlexG - Gender Inflexion Pointer

VALUES: *v* = gender inflexion permissible; *vs* = gender inflexion possible, but biased by socio-cultural pragmatics; [empty] if gender inflexion not permissible.

CONCEPT: Code for whether gender inflexion of nouns, adjectives, and numerals is permissible or not. The special case where gender inflexion is typically not used, or only very rarely, due to gender related socio-cultural stereotypes is separately marked. For example, the feminine form for *pedreiro*, mason, though grammatically acceptable is practically never used because there are hardly any women-masons.

21. Plur - Plural Pointer

VALUES: *pl* = plural; [empty] if singular.

CONCEPT: Code for the classification of nouns, adjectives, numerals and pronouns according to grammatical number: plural or singular. Since almost all entries are in the singular form (because they are uninflected), only the plural form is marked.

22. FreqL - Lexical Frequency

VALUES: *1, 2, ... 34 740;* [empty] if no frequency information is available.

CONCEPT: Frequency of the word as computed by Nascimento et al. (1987a, 1987b) from a corpus of 700 thousand wordtokens of spoken European Portuguese. The value refers to the number of occurrences of inflected wordtypes or of lemmas, according to type of Porlex entry (function or content word). Frequency information is available for 5% of the words in Porlex.

23. DO - Orthographic Density

VALUES: *0, 1, . . ., 37.*

CONCEPT: Number of orthographic neighbors, where neighbor is any other orthographical entry that differs on one segment only, by substitution or by addition/deletion, while preserving relative position of the segments. This value was computed by an automatic serial check of all orthographic entries, excluding homographs, and was sensitive to diacritics. E.g., *sala* has 15 neighbors (cf. infra).

24. VO - Orthographic Neighbors

VALUES: *ala, bala, fala, gala, mala, pala, saca, saga, saia, sal, salsa, sela, sola, tala, vala;* …; [empty] if no neighbors.

CONCEPT: List of orthographic neighbors, alphabetically ordered.

25. PUO - Orthographic Uniqueness Point

VALUES: *1, 2, . . ., 20.*

CONCEPT: Position of the letter, counted from the start, that uniquely identifies the orthographic wordform from its neighbors. E.g., for *sala* PU = 4 because the preceding letters are shared with *sal*.

26. DFot - Phonetic Density

VALUES: *0, 1, . . ., 25.*

CONCEPT: Number of phonetic neighbors, where neighbor is any other phonetic entry that differs on one segment only, by substitution or by addition/deletion, while preserving relative position of the segments. Homophones were excluded from the computation. This value was computed by an automatic serial check of all phonetic entries, excluding homophones, and was not sensitive to stress. E.g., *salA* has 14 phonetic neighbors (cf. infra).

27. VFot - Phonetic Neighbors

VALUES: *alA, balA, falA, galA, malA, palA, sakA, saGA, sajA, sElA, silA, sOlA, talA, valA*; …; [empty] if no neighbors.

CONCEPT: List of phonetic neighbors of the wordform.

28. PUFot - Phonetic Uniqueness Point

VALUES: *1, 2, . . ., 19.*

CONCEPT: Position of the phone, counted from the start, that uniquely identifies the phonetic wordform from its neighbors. E.g., for *salA* PU = 3 because the preceding phones are shared with *saKA* (note that due to the velarization of the final /l/, the phonetic transcription of the word *sal* also differs in the third phone, *sa9*).

29. DFom - Phonemic Density

VALUES: *0, 1, . . ., 28.*

CONCEPT: Number of phonemic neighbors, where neighbor is any other phonemic entry that differs on one segment only, by substitution or by addition/deletion, while preserving relative position of the segments. This value was computed by an automatic serial check of all phonemic entries, excluding homophones, and was not sensitive to stress. E.g., *salA* has 16 phonemic neighbors (cf. infra).

30. VFom - Phonemic Neighbors

VALUES: *alA, balA, falA, gala, malA, palA, sakA, saga, sajA, sal, salsA, sElA, silA, sOlA, talA, valA*; …; [empty] if no neighbors.

CONCEPT: List of phonemic neighbors of a given wordform.

31. PUFom - Phonemic Uniqueness Point

VALUES: *1, 2, . . ., 18.*

CONCEPT: Position of the phoneme, counted from the start, that uniquely identifies the phonemic wordform from its neighbors. E.g., for *salA* PU = 4 because the preceding phones are shared with *sal*.

32. HGnF - Nonhomophonic Homographs

VALUES: *1, 2, 3*; [empty] if no nonhomophonic homographs.

CONCEPT: For a given orthographic wordform, number of identical entries that differ on phonetic/phonemic wordform. This is one of the two sole instances of repeated orthographic wordforms (for the other, see below). E.g., *colher*, [ku'LEr], and *colher*, to [ku'Ler]. Since these cases are scarce, only positive cases are marked.

33. HGHF - Homographic Homophones

VALUES: *1, 2, 3, 4*; [empty] if no homographic homophones.

CONCEPT: Number of entries with identical orthographic and phonetic/phonemic wordforms that differ on grammatical class. This follows from the criterion adopted to establish the database, and it is one of the two sole instances of repeated entries (see above). Since homographic homophones are scarce, only positive cases are marked.

34. HFnG - Nonhomographic Homophones

VALUES: *1, 2, 3, 4,* [empty] if no nonhomographic homophones.

CONCEPT: For a given orthographic wordform, number of entries with identical phonetic/phonemic sequences irrespective of stress position. Strictly speaking, these are segmental homophones; e.g., *túnel,* [‘tunE9] and *tonel,* [tu’nE9]. Since these cases are scarce, only positive cases are marked.

35. PFot1 - Gross Phonetic Pattern

VALUES: V’CV.CVC, VC.CVC’CV.CV, . . .; where *C* = consonant, *V* = vowel, *G* = semivowel, and H = homorganic nasal.

CONCEPT: Classification of the segments of the phonetic wordform into consonant, vowel or glide types (C, V, G, respectively). Because of the coarticulatory nature of homorganic nasals, these were classified separately (H). Syllable boundaries and stress marks are shown.

36. PFot2 - Detailed Phonetic Pattern

VALUES: V’SV.ZVL, VL.PVF’PV.PV, . . .; where *P* = voiceless stop; *B* = voiced stop; *S* = voiceless fricative; *Z* = voiced fricative; *L* = lateral approximant; *N* = nasal; *V* = oral vowel; *M* = nasal vowel; *G* = oral semivowel; *W* = nasal semivowel; *R* = trill; *F* = flap; *D* = fricatization of /b, d, g/; *H* = homorganic nasal.

CONCEPT: Classification of the segments of the phonetic wordform according to manner (stop, nasal, trill, flap, fricative, approximant; vowel), voicing, and oral vs. nasal quality. Coarticulatory homorganic nasals were included as a separate category. Syllable boundaries and stress marks are shown.

37. InvO - Reverse Orthographic Wordform

VALUES: *levàfa, atacrepla, acenob, ocitírc, ahnel, ogsum, alas, . . .*

CONCEPT: Backward sequence of the orthographic wordform (from variable 2, Orthographic Wordform, thus with diactrics).

38. InvF - Reverse Phonetic Wordform

VALUES: *9EvafA, Atakr6p9a, AkEnub, ukitirk, ANAl, ugZum, Alas, . . .*

CONCEPT: Backward sequence of the phonetic wordform.

39. Maius - Uppercase Wordform

VALUES: *AFAVEL, ALPERCATA, BONECA, CRITICO, LENHA, MUSGO, SALA*, . . .

CONCEPT: Orthographic wordform in uppercase without diacritics (but with cedilla).

40. VFot1 - Phonetic Variant 1

41. VFot2 - Phonetic Variant 2

42. VFot3 - Phonetic Variant 3

VALUES: *ˈleNA, ˈlAjNA, ˈlENA;* …, [empty] if no phonetic variant.

CONCEPT: Alternative phonetic transcription(s) of the wordform, [empty] if no phonetic variants or if there are less than 3 phonetic variants.

43. VarLex - Lexical Variant Orthographic Wordform

VALUES: *alparcata, alpergata, alpargata; bonecra*; . . .; [empty] if no lexical variant.

CONCEPT: Orthographic wordform of the lexical variant (lower case with diacritics).

44. FotVarL - Lexical Variant Phonetic Wordform

VALUES: *a9.pAr'ka.tA, a9.p6r'ga.tA; a9.pAr'ga.tA; buˈnEkrA;* . . .; [empty] if no lexical variant.

CONCEPT: Phonetic transcription of the lexical variant.

Acknowledgments

The authors would like to acknowledge Porto Editora for making available the electronic version of the entries from Dicionário da Lingua Portuguesa (Costa, & Melo, 1997). This research was supported in part by a FCT grant to the Center for Psychology at the University of Porto (Language Group). I. Gomes was supported by a FCT grant (PRAXIS XXI/BD/4529/94).

Correspondence should be addressed to: São Luís Castro, FPCE-Universidade do Porto, rua Campo Alegre, 1021, P 4169 - 004 Porto, Portugal (slcastro@psi.up.pt).

Footnotes

1. Porlex was started in 1998, and the final computations were completed in early 2000. DLPC, the first Portuguese dictionary to provide phonetic transcriptions in European Portuguese for a middle sized vocabulary, appeared in early 2001 (Academia das Ciências de Lisboa, 2001). The entries from the Dicionário da Língua Portuguesa (Costa & Melo, 1997) were made available to us in electronic format, but they had to be checked individually in order to insert source information on phonological characteristics and syllabication, that we were unable to find in a compatible electronic format. CORLEX, a corpus that comprises 26 443 lemmas and 140 315 wordforms, as well frequency information based on 16 210 438 wordtokens, is now available at the Centro de Linguística da Universidade de Lisboa website at http://www.clul.ul.pt/sectores/projecto lmcpc.html.

References

A. Sources for Porlex

Costa, J. A., & Melo, A. S. (1997). *Dicionário da língua portuguesa* [Portuguese language dictionary] (7th Ed.). Porto: Porto Editora.

Cunha, C., & Cintra, L. F. L. (1987). *Nova gramática do português contemporâneo* [New grammar of contemporary Portuguese] (4ª Ed.). Lisboa: Edições João Sá da Costa.

Figueiredo, C. (1996). *Grande dicionário da língua portuguesa* [Extended Portuguese language dictionary] (25th Ed.). Venda Nova: Bertrand Editora.

Irmen, F., & Kollert, A. M. C. (1995). *Langenscheidts Taschenwörterbuch Portugiesisch* [Langenscheidt Portuguese pocket dictionary]. Munich: Langenscheidt.

Mateus, M. H. M., Brito, A. M., Duarte, I., & Faria, I. H. (1989). *Gramática da língua portuguesa* [Portuguese language grammar] (4ª Ed.). Lisboa: Editorial Caminho.

Melhoramentos (1998). *Michaelis: Pequeno dicionário da língua portuguesa* [Michaelis: Brief Portuguese dictionary]. São Paulo: Author.

Nascimento, M. F. B., Marques, M. L. G., & Cruz, M. L. S. (1987). *Português fundamental: Métodos e documentos* (Vol. II, Tomo I: Inquérito de frequência) [Basic Portuguese: Methods and documents. Vol. II, Tomo I: Frequency survey]. Lisbon: INIC, Centro de Linguística da Universidade de Lisboa.

Nascimento, M. F. B., Rivenc, P., & Cruz, M. L. S. (1987). *Português fundamental: Métodos e documentos* (Vol. II, Tomo II: Inquérito de disponibilidade) [Basic Portuguese: Methods and documents. Vol. II, Tomo II: Availability survey]. Lisbon: INIC, Centro de Linguística da Universidade de Lisboa.

Texto Editora (1998). *Dicionário universal fundamental da língua portuguesa* [Basic universal Portuguese language dictionary] (1st Ed.). Lisbon: Author.

Vilela, M. (1991). *Dicionário do português básico* [Elementary Portuguese dictionary] (3rd Ed.). Rio Tinto: Edições Asa.

Vilela, M. (1995). *Gramática da língua portuguesa* [Portuguese language grammar]. Coimbra: Livraria Almedina.

fine tune

B. Others

Academia das Ciências de Lisboa (2001). *Dicionário da Língua Portuguesa Contemporânea (DLPC)* [Dictionary of contemporary Portuguese language]. Lisboa: Editorial Verbo.

Azuaga, L. (1996). Morfologia [Morphology]. In I. H. Faria, E. R. Pedro, I. Duarte, & C. A. M. Gouveia (Eds.), *Introdução à linguística geral e portuguesa* [An introduction to general and Portuguese linguistics] (pp. 215-244). Lisbon: Editorial Caminho.

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

Burnage, G. (1990). *CELEX – A guide for users.* Nijmegen: Centre for Lexical Information, University of Nijmegen.

Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children's lexicons. *Journal of Child Language, 17*, 205-215.

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

Content, A., Mousty, P., & Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le Français écrit et parlé [Brulex: A computerized lexical database for written and spoken French]. *L'Année Psychologique, 90,* 551-566.

Fromkin, V., & Rodman, R. (1998). *An introduction to language* (6th Ed.). Fort Worth: Harcourt Brace.

Gomes, I. (2001). *Ler e escrever em Português Europeu* [Reading and writing in European Portuguese]. Unpublished doctoral dissertation, University of Porto, Faculdade de Psicologia e de Ciências da Educação, Portugal.

Lambert, É., & Chesnet, D. (2001). Novlex: Une base de données lexicales pour les élèves de primaire [Novlex: A lexical database for elementary school students]. *L'Année Psychologique, 101,* 277-288.

Luce, P. A. (1986). *Neighborhoods of words in the mental lexicon* (Tech. Rep. No. 6). Bloomington, Indiana: University of Indiana, Speech Research Laboratory, Department of Psychology.

Marslen-Wilson, W. D. (1990). Activation, competition, and frequency in lexical acess. In G. T. Altman (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148-172). Cambridge: Bradford Books.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8*, 1-71.

Microsoft Corporation (1998). *Microsoft Office 98 – Macintosh Edition* [Computer software]. USA: Author.

Nascimento, M. F. B., Casteleiro, J. M., Marques, M. L. G., Barreto, F., & Amaro, R. (n.d.). Léxico multifuncional computorizado do Português Contemporâneo [Multifunctional computacional lexicon of contemporary Portuguese]. Available: http://www.clul.ul.pt/sectores/projecto lmcpc.html/ [2002, Dec. 13].

Nascimento, M. F. B., Rodrigues, M. C., & Gonçalves, J. B. (Eds.). (1996). *Actas do XI Encontro Nacional da Associação Portuguesa de Linguística* (Vol. I – Corpora)

[Proceedings of the XI National Meeting of the Portuguese Linguistic Association]. Lisbon: Colibri – Artes Gráficas.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du Français contemporain sur internet: Lexique^{TM} [A lexical database for contemporary French on Internet: Lexique]. *L'Année Psychologique, 101*, 447-462.

Piepenbrock, R. (2001, Feb. 18). *Celex, The Dutch centre for lexical information* [On line]. Available: http://www.kun.nl/celex/ [2002, Nov. 26].

Piñeiro, A., & Manzano, M. (2000). A lexical database for Spanish-speaking children. *Behavior Research Methods, Instruments, & Computers, 32,* 616-628.

Segalowitz, S. J., & Lane, K. C. (2000). Lexical access of function versus content words. *Brain and Language, 75,* 376-389.

Table 1. Information available in Porlex for each word

Variable #[1]	Code Name	Type[2]	Content
1	#	I	Number
2*	Orto	S	Orthographic Wordform
3	Diacr	C	Diacritic Pointer
4*	Fot	S	Phonetic Wordform
5	Fom	S	Phonemic Wordform
6*	Var	C	Variant Pointer
7*	CGram	C	Grammatical Class
8	CAF	C	Open/Closed Class
9	Tonic	C	Stress Position
10	Nlet	I	Letter Length
11	NSilO	I	Orthographic Syllable Length
12	Nfom	I	Phonemic Length
13	NHC	I	N Homorganic Nasals
14	NSilF	I	Phonological Syllable Length
15	NSchw	I	N Schwa
16*	DivSilO	S	Orthographic Syllabication
17*	DivSilF	S	Phonological Syllabication
18*	AmbSil	C	Ambisyllabicity Pointer
19*	Gen	C	Gender
20*	FlexG	C	Gender Inflexion Pointer
21*	Plur	C	Plural Pointer
22*	FreqL	I	Lexical Frequency
23r	DO	I	Orthographic Density
24r	VO	S	Orthographic Neighbors
25r	PUO	I	Orthographic Uniqueness Point
26r	DFot	I	Phonetic Density
27r	VFot	S	Phonetic Neighbors
28r	PUFot	I	Phonetic Uniqueness Point
29r	DFom	I	Phonemic Density
30r	VFom	S	Phonemic Neighbors
31r	PUFom	I	Phonemic Uniqueness Point
32r	HGnF	I	Nonhomophonic Homographs
33r	HGHF	I	Homographic Homophones
34r	HFnG	I	Nonhomographic Homophones
35	PFot1	C, S	Gross Phonetic Pattern
36	PFot2	C, S	Detailed Phonetic Pattern
37	InvO	S	Reverse Orthographic Wordform
38	InvF	S	Reverse Phonetic Wordform
39	Maius	S	Uppercase Wordform
40*	VFot1	S	Phonetic Variant 1
41*	VFot2	S	Phonetic Variant 2
42*	VFot3	S	Phonetic Variant 3
43*	VLex	S	Lexical Variant Orthographic Wordform
44*	FotVarL	S	Lexical Variant Phonetic Wordform

Note. [1]The asterisc is used to indicate source information. Subscript r indicates relational computations.

[2]Entry type: C = category; I = integer; S = string

Table 2. Number and percentage of lexical entries by word class, split by gender

			Gender		
Word Class	*n*	%	Masculine	Feminine	Invariant
Noun	16,313	55.78	7,952	7,707	654
Adjective	5,806	19.86	3,826	17	1,963
Verb	5,508	18.84	-	-	-
Adverb	1,056	3.61	-	-	-
Preposition	21	0.07	-	-	-
Article	8	0.03	4	4	-
Pronoun	147	0.50	57	52	32
Conjunction	33	0.11	-	-	-
Contraction	118	0.40	54	50	5
Numeral	55	0.19	29	-	26
Interjection	59	0.20	-	-	-
Locution	113	0.39	-	-	-

Appendix A

Table A1. Overview of corpora in European Portuguese I (as described in Nascimento, Rodrigues & Gonçalves, 1996).

#	Name (Author & Afiliation[1])	Number of wordforms[2]	Annotation (or Authors description)
1	Corpus Referência Português Contemporâneo (Casteleiro & Nascimento, CLUL)***	45 millions	Morphosyntactic and syntactic codes
2	NATURA-PÚBLICO (Almeida, UM)*	ca 6 milions	None
3	NATURA-PÚBLICO-Etiquetado (Almeida, UM)*	ca 4,000	Morphosyntactic class
4	ONOMASTICA (Trancoso, INESC, CLUL)**	100,000 proper names, siglas, acronyms	Broad phonetic transcription inc. sillabication
5	PF-FONE (Viana & D'Andrade, CLUL)**	26,000	Phonetic transcription inc. sillabication
6	COPUSINESC (Santos, INESC)*	14,873 (in 1,000 sentences)	Word class
7	EUROM.1 Português (Trancoso, INESC, CLUL, PT)**	6,500 numbers, 2,200 sentences, 1,260 words	Broad phonetic transcription
8	CIPM – Português Medieval (DEL team, FCSH – UNL)*	154,122; 244,775; 223,095 (XIII to XV century)	Morphosyntactic class
9	Textos metalinguísticos portugueses do século XVI (Paiva, FLUP)*	65,730	(Linguistic and metalinguistic codes)
10	Moda 60-90 (Carvalho)*	700,000	–
11	Astro (Neto, CLUL)**	560,000	–
12	RED-I (Andrade, CLUL)**	–	(Words and sentences)
13	PROPER-(PE) (Andrade, CLUL)**	–	(Speech material and perceptive data)
14	BDFALA (Trancoso & Viana, INESC, CLUL)**	–	(Lexemes, words, sentences, and texts from 10 speakers)

Note. The asterisk indicates whether source materials are written only*, spoken only** or both***. The sign – indicates that the information was not available.

[1] UM = Universidade do Minho; INESC = Instituto Nacional de Engenharia; DEL = Departamento de Estudos Linguísticos; FCSH-UNL = Faculdade de Ciências Sociais e Humanas, Universidade Nova de Lisboa; FLUP = Faculdade de Letras, Universidade do Porto; CLUL = Centro de Linguística, Universidade de Lisboa; PT = Portugal Telecom.

[2] This includes repetitions and inflected forms, unless otherwise stated.

Table A2. Overview of corpora in European Portuguese II: Speech in interviews or interactions (ibd.)

#	Name (Author &	Content	Annotation (or

	Afiliation[1])		Authors description)
1	100 Falantes Adultos do PE (Faria, FLUL)*	Structured interviews	CLAN
2	Batoréo 94 (Batoréo, FLUL)*	Narratives from adults and children (60 + 60)	CLAN
3	Subcorpus interacções mãe / criança corpus aquisição PE (Ramos & Faria, FLUL)*	30 dyads mother-child	CLAN
4	Corpus aquisição PE (Fonologia) (Faria & Freitas, FLUL)*	Spontaneous speech (8 children aged 0;11 to 3;07 years-old)	Broad phonetic transcription
5	CPE FACES – *Corpus* PE Falado Adolescentes contexto escolar (Mata, FLUL)*	Spoken texts (15 hours of tape recording)	Prosodic marks
6	Atlas Linguarum Europae (Barros, CLUL)*	Lexical questionary (546 question-tags)	Phonetic transcription
7	Atlas Linguístico Portugal e Galiza (Dialectology team, CLUL)*	Linguistic questionary	Phonetic transcription
8	Atlas Linguístico Litoral Português (Vitorino, CLUL)*	Linguistic questionary	Phonetic transcription
9	SPEECHDAT(Português) (Trancoso, PT)*	Phone calls	None
10	Panorama Português oral Maputo (Gonçalves & Stroud, INDE)*	50 interviews	None
11	CPE VAR (D'Andrade e Rodrigues, FLUL/CLUL)**	30 interviews and material from reading	–
12	Leiria (Leiria, FLUL)*	218 narratives	–

Nota. The asterisk indicates whether source materials are spoken only* or written and spoken**. The sign – indicates that the information was not available. PE = European Portuguese.

[1] INDE = Instituto Nacional de Desenvolvimento da Educação; FLUL = Faculdade de Letras, Universidade de Lisboa; CLUL = Centro de Linguística, Universidade de Lisboa; PT = Portugal Telecom;

Appendix B
List of the phonetic symbols used in Porlex

These phonetic symbols are an adaptation of the Unibet system for European Portuguese (Castro, S. L., & Gomes, I. [2001]. *O sistema Unibet adaptado ao Português Europeu.* [Laboratório de Fala, FPCE-UP.]; cf. MacWhinney, B. [1995]. *The Childes project: Tools for analyzing talk* [pp. 67-79]. Hillsdale, N.J.: LEA).

Unibet	AFI	Exemplo	Unibet	AFI	Example
a	a	s**a**co	t	t	**t**apete
A	♊	c**a**m**a**	d	d	**d**ado
E	ε	f**e**rro	D	δ	í**d**olo
e	e	s**e**de	k	k	**c**acto
6	☜📪 ✋	m**e**lão	g	g	**g**ato
i	i	l**i**vro	G	📪	á**g**ua
O	✂	comb**o**io	f	f	**f**aca
o	o	av**ô**	v	v	**v**ela
u	u	**u**va	s	s	**s**ola
1	⮰	d**an**ça	z	z	**z**ebra
2	🡅	p**en**te	S	🗁	**x**adrez
3	⬋	c**in**zeiro	Z	☯	**g**elado
4	õ	p**on**te	m	m	**m**açã
5	▫	m**un**do	n	n	**n**ariz
j	j	pa**i**	N	✞	ni**nh**o
w	w	pa**u**	K	☠	z**an**ga
7	🡄	mã**e**	l	l	**l**imão
8	⑦	mã**o**	9	⌨	me**l**
p	p	**p**ato	L	✡	mi**lh**o
b	b	**b**atata	r	🗎	ca**r**o
B	β	tá**b**ua	R	R, 🕯	**r**ato