Porlex

1. BASIC INFORMATION

1.1 *Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc.)*

Porlex v2 (see Gomes & Castro, 2003) is a computerized lexical database in European Portuguese containing psycholinguistic and cognitive information that is useful to select stimulus materials for experiments and/or training vocabularies. It was built on the basis of a middle-sized adult lexicon, and provides orthographic, phonological, phonetic, part-of-speech, and neighborhood information for about 27,374 words (uninflected content words and inflected function words), presented as an Excel file.

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).

Lexical entries, grammatical and morphological classification, syllabication, phonetic transcription and frequency information (~5% of the lexical entries, imported from the Fundamental Portuguese – a spoken corpus) were collected from various sources. Porlex gives additional information 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 this and more information, please visit the Porlex webpage: http://www.fpce.up.pt/labfala/research.html.

1.2 *Representation of the lexicon (flat files, database, markup)*

The corpus is represented in an Excel format.

1.3 *Character encoding*

The characters are in UTF8 encoding.

2. ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

Name: São Luís Castro
Address:
Laboratório de Fala da Faculdade de Psicologia e de Ciências da Educação, Sala FPCE 010A
Rua Doutor Manuel Pereira da Silva 4200-392 Porto
Position: Professor
Affiliation: Laboratório de Fala da Faculdade de Psicologia e de Ciências da Educação – Universidade do Porto
Telephone: +351 220 400 610
Fax: -
E-mail: labfala@fpce.up.pt

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource is available on the META-SHARE platform.

2.3 Copyright statement and information on IPR

This resource is free for research purposes, with attribution and no redistribution allowed. It is available on the META-SHARE platform only upon request to the authors.

3. TECHNICAL INFORMATION

3.1 Directories and files

The archive that can be downloaded on the META-SHARE is a .zip file with X files...

3.2 Data structure of an entry

Porlex entries are organized in alphabetical order. Each lexical entry comprises all forty four information fields (see table in 4.3). Each entry is an excel line for which each cell correspond to one information field. To better demonstrate that, see the following example in CSV format for the entry `talento` (talent):

```
#,Orto,Diacr,Fot,Fom,Var,CGram,CAF,Tonic,Nlet,NSilO,Nfom,NHC,NSilF,NSchwa,DivSilO,DivSilF,AmbSil,Gen,FlexG,Plur,FreqL,DO,VO,PUO,DFot,VFot,PUFot,DFom,VFom,PUFom,HGnF,HGHF,HFnG,PFot1,PFot2,InvO,InvF,Maiusc,VFot1,VFot2,VFot3,VarLex,FotVarL,,

25363,talento,0,tA`l2ntu,tA`l2tu,,no,a,gr,7,3,6,1,3,,ta-len-to,tA`l2n.tu,,m,,,,1,alento,7,1,Al2ntu,7,1,Al2tu,6,,,,CV`CVH.CV,PV`LMH.PV,otnelat,utn2lAt,TALENTO,,,,,,,
```

3.3 Lexicon size (nmb. of lexical items, KB occupied on disk)

The lexicon is composed by 27,374 words with 14,3 MB of disk storage.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

The language of Porlex is European Portuguese.

4.2 Entry Type

For this information, please see item 3.2.

4.3 Attributes and their values

For attributes and values of Porlex, please the following table:

Information available in Porlex for each word			
Variable #1	**Code Name**	**Type2**	**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

1. The asterisk is used to indicate source information. Subscript r indicates relational computations.

2. Entry type:

C = category
I = integer
S = string

4.4 Coverage of the lexicon

Porlex covers the general language of a middle-sized adult lexicon.

4.5 *Intended application of the lexicon*

Since this dataset contains psycholinguistic and cognitive information that is useful to select stimulus materials for experiments and/or training vocabularies, Porlex has been used for experimental research on language, mainly in psycholinguistic field of study.

4.6 POS assignment

Not applicable.

4.7 Reliability (automatically/manually constructed)

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 were 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.

5. RELEVANT REFERENCES AND OTHER INFORMATION

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