



CELGA-ILTEC

Centro de Estudos de Linguística Geral
e Aplicada da Universidade de Coimbra

U LISBOA

UNIVERSIDADE
DE LISBOA



VOC, a Spelling Dictionary for the Portuguese Language

– role and characteristics

Margarita Correia

CELGA-ILTEC e FLUL

eLex 2019

Sintra, October 1st 2019




Structure


- **VOC – some history**
 - Portuguese language orthography
 - Outset and function
- **Some lexicographic challenges and issues**
 - Single but multiple
 - Dealing with variation
- **Political role and impact**
 - Pluricentric vision of the language
 - Management of Portuguese language orthography

VOC – some history

- The spelling of the Portuguese language is set in a legally binding document to which all official documents have to adhere.
- For the whole 20th century and until recently Portuguese orthography was dealt with in a bicentric setting – BR / PT (Marquilhas 2015)
- In 1990, a common document was finally agreed upon, called the Acordo Ortográfico da Língua Portuguesa (AOLP90).

- AOLP90 consists of a set of rules defining how to write the words of Portuguese. An official part of the agreement is that the practical implementation of those rules is detailed in an official spelling dictionary, called the Vocabulário Ortográfico Comum da Língua Portuguesa (Common Spelling Dictionary of the Portuguese Language – VOC)










- 
- VOC is the practical implementation of the AOLP90 spelling rules, defining explicitly, based on those rules, what the orthography of concrete words should be for Portuguese.
 - VOC is organized under the guidance of the International Institute for the Portuguese Language (IILP), a body of the Comunidade de Países da Língua Portuguesa (CPLP - Community of Portuguese Speaking Countries), which officially recognized and published it in 2017.

- 
- VOC is a common spelling dictionary valid in all Portuguese-speaking countries.
 - AOLP90 allows for some national-level spelling variation:
 - the CPLP countries introduced the notion of national spelling dictionary (VON – Vocabulário Ortográfico Nacional), containing the nationally-representative words and, in some instances, national-level variants.

- The VOC lexicon is officially hosted at the site of the CPLP, but also included in other sites, most prominently the Portal da Língua Portuguesa.
- The default way in which those interfaces behave is to guess the preference of the user (based on locale and country of the request IP) or ask for it explicitly, and display the spelling recommendation specific to the selected VON.
- However, it is also possible to have access to the entire VOC lexicon.

Vocabulário Ortográfico Comum da Língua Portuguesa

Selecione a versão do VOC a usar

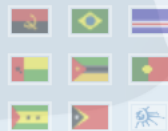
 Angola	ainda não disponível
 Brasil	VOLP: Vocabulário Ortográfico da Língua Portuguesa
 Cabo Verde	VOCALP: Vocabulário Cabo-Verdiano da Língua Portuguesa
 Guiné-Bissau	ainda não disponível
 Moçambique	VONMoz: Vocabulário Ortográfico Nacional de Moçambique
 Portugal	VOP: Vocabulário Ortográfico do Português
 São Tomé e Príncipe	ainda não disponível
 Timor-Leste	VO-TL: Vocabulário Ortográfico de Timor-Leste
 todos os países	versão comum

O VOC tem versões específicas para cada país, refletindo as fontes, a frequência e as propriedades das formas mais representativas de cada país. Clique na bandeira acima para selecionar a versão que pretende usar. Pode alterar essa definição a qualquer momento através das bandeiras na coluna da esquerda.

De modo a permitir esta funcionalidade, o VOC guarda alguma informação (*cookie*) para cada utilizador. Ao continuar a navegar está a consentir a sua utilização.

igual a

Busca



Apresentação

Vocabulário


Toponímia


Formas não adaptadas

Ficha técnica

Some lexicographic challenges and issues

- VOC contains over 300k lexical-entries, distributed by three areas:
 - general lexicon;
 - loan words;
 - toponyms;

- 
- AOLP90 explicitly allows for some degree of variation in spelling, reflecting in some orthographic contexts the way words are pronounced differently in different countries.
 - Each VON specifies:
 - which of the words in VOC are considered a core part of each national vocabulary,
 - for those words that allow spelling variation, which of the spelling options are recommendable for that country.

- 
- Each VON is in principle based on two pillars:
 - the lexicographic memory of the country, that is, the words included in the reference dictionarie(s) of the country,
 - the frequency of each word in a corpus of the national variety, the CPLP corpus itself a pluricentric corpus, created for the purpose of VOC.

Vocabulário Ortográfico Comum da Língua Portuguesa



capulana - feminino

(ca.pu.'la.na)

singular capulana
plural capulanas

Fontes

Corpus Brasileiro: **média**

Corpus Moçambicano: **alta**

Corpus Português: **baixa**



Vocabulário Ortográfico Comum da Língua Portuguesa

igual a Busca




Apresentação

Vocabulário

Toponímia

Formas não adaptadas

Ficha técnica

- 
- VOC is implemented in a lexical management platform called the Open Source Lexical Information Network (OSLIN) – Janssen, 2005 –, designed as a relational database, with tables for basic lexicographic notions and relations between them.
 - Each lexical entry contains a citation form and a word-class, as well as additional information such as the syllabic structure of the word and its inflectional paradigm.

Vocabulário Ortográfico Comum da Língua Portuguesa

VOP: Vocabulário Ortográfico do Português, 2.ª edição

falar - verbo

(fa.'lar)

Indicativo

	Presente	Pretérito imperfeito	Pretérito perfeito
eu	falo	falava	falei
tu	falas	falavas	falaste
ele/ela	fala	falava	falou
nós	falamos	falávamos	falámos / falamos
vós	falais	faláveis	falastes
eles/elas	falam	falavam	falaram

Pretérito mais-que-perfeito

eu	falara
tu	falaras
ele/ela	falara
nós	faláramos
vós	faláreis
eles/elas	falaram

Futuro imperfeito

eu	falarei
tu	falarás
ele/ela	falará
nós	falaremos
vós	falareis
eles/elas	falarão

Futuro perfeito (condicional)

eu	falaria
tu	falarias
ele/ela	falaria
nós	falaríamos
vós	falaríeis
eles/elas	falariam

Conjuntivo / Subjuntivo

	Presente	Pretérito imperfeito	Futuro
eu	fale	falasse	falar
tu	fales	falasses	falares
ele/ela	fale	falasse	falar
nós	falemos	falássemos	falarmos
vós	faleis	falásseis	falardes
eles/elas	falem	falassem	falarem

Otras Formas

	Imperativo afirmativo (negativo)	Infinitivo flexionado	Formas nominais
eu		falar	Infinitivo falar Gerúndio falando
tu	fala (fales)	falares	
ele/ela	fale	falar	Particípio passado falado
nós	falemos	falarmos	
vós	falai (faleis)	falardes	
eles/elas	falem	falarem	

► ver fontes



Vocabulário Ortográfico Comum da Língua Portuguesa

igual a Busca




[Apresentação](#)

[Vocabulário](#)

[Toponímia](#)

[Formas não adaptadas](#)

[Ficha técnica](#)

- 
- When considering a national variety, there is scale of acceptability: from words that are in the official VON, to forms that are deemed not acceptable in that variety.
 - In its core VOC consists of an index table with nine columns: the first indicating the ID of the word, and the other columns the acceptability index for each of the eight countries participating in the project. At the database level, there are no separate entry lists for each country – Janssen; Ferreira 2018.



1 A word explicitly registered in a primary source for the VON

2 A word of unrestricted use in the country, but not explicitly registered in a primary source


3 A word not recommendable in the VON due to usage considerations


4 A word not representative of the country's lexicon due to country-specific AOLP90 variation choices


5 A word fully unacceptable in the VON


Political role and impact

- VOC constitutes a milestone for Portuguese language (electronic) lexicography.
- Portuguese language is now provided with an electronic large scale spelling dictionary, which can become the basis for:
 - new derived lexicographic resources;
 - NLP tools involving lexical knowledge.
- VOC has been conceived as a permanent work in progress

- 
- VOC is part of a new political perception of the Portuguese as a pluricentric language, with several emerging national standards besides those of Brazil and Portugal.
 - All previous lexicographic products, including spelling dictionaries, were nationally oriented and adopted a contrastive stance. VOC is a real pluricentric dictionary for the Portuguese language.
 - VOC is not a contrastive dictionary, but rather integral. (Lara)

- 
- VOC inaugurates a new perspective on the international management of Portuguese:
 - IILP, where all CPLP member-states are represented as peers, and which incorporates the first official resources of the Portuguese language for a significant number of countries.
 - VOC intends to launch a shared management of the Portuguese language orthography, in which all countries have equal status.

- 
- As a consequence of VOC, IILP created the Council for the Portuguese Language Orthography (COLP).
 - All decisions on Portuguese language spelling will be made in a diplomatic context by official representatives of all Portuguese speaking countries.
 - VOC allows for the end to the hundred-year long “orthographic schism” started in the early 20th century.

- 
- VOC is thus
 - - a lexicographic resource,
 - a significant asset for linguistic policy.

Bibliography

- Buchmann, F. 2016. Spelling dictionaries. In Durkin, P., ed.: *The Oxford Handbook of Lexicography*. Oxford University Press
- Ferreira, J.P., Correia, M., de Almeida, G.B., eds.: 2017. *Vocabulário Ortográfico Comum da Língua Portuguesa*. Instituto Internacional da Língua Portuguesa / Comunidade dos Países de Língua Portuguesa, Praia, Cape Verde / Lisbon, Portugal
- Ferreira, J.P., Janssen, M., de Almeida, G.B., Correia, M., de Oliveira, F.M. 2012. The Common Orthographic Vocabulary of the Portuguese language: a set of open lexical resources for a pluricentric language. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 1071–1075
- Janssen, M.: 2005. Open source lexical information network. In: *3rd International Workshop on Generative Approaches to the Lexicon*, Geneva,
- Janssen M., Ferreira J.P. 2018. Technical Implementation of the Vocabulário Ortográfico Comum da Língua Portuguesa. In: Villavicencio A. et al. (eds) *Computational Processing of the Portuguese Language. PROPOR 2018. Lecture Notes in Computer Science*, vol 11122. Springer, Cham.
- Janssen, M., Kuhn, T.Z., Ferreira, J.P., Correia, M. 2018. The CPLP corpus, a corpus of Portuguese as a pluricentric language. In: *XVIII EURALEX International Congress*, Ljubljana, Slovenia
- Lara, L. F. 1997. *Teoría del diccionario monolingüe*. México: El Colegio de México.
- Marquilhas, R. 2015. The Portuguese language spelling accord. *Written Language Literacy* 18(2) 275–286



Thank you.

margarita@campus.ul.pt