Using wiktionary for advances on Portuguese phonology

Luís Trigo[*]
LIACC Laboratório de Inteligência Artificial e Ciência de Computadores, Faculty of Engineering University of Porto
Porto, Portugal

Carlos Silva
CLUP Centro de Linguística, University of Porto, Porto, Portugal
Porto, Portugal

ABSTRACT

As a language-related project, Wiktionary is a powerful tool for linguistic research [5]. In spite of the large amount of language data stored online, open-source corpora of phonetically transcribed data are rare, which makes Wiktionary crucial to phonological studies in particular [7]. Phonology can use such corpus in, at least two dimensions:

(1) to evaluate the behavior of a given phoneme, e.g. overall frequency, preferred adjacency and word position;
(2) to assess the behavior of syllables, e.g. frequency of syllable templates, stress;

These approaches allow great progress and can have wide range of applications, from theoretical phonology to the clinical treatment of speech disorders. Moreover, it can be used for language comparison. However, to enable them, one should depart from a clean and reliable data set.

The wikcionário version collected in from the dump repository in January, 18th 2022 had 12,372 with IPA transcription. By doing some visual inspection and using basic filters, we found an error rate of about 3% on the IPA transcription and syllabification in Wikcionário (the Portuguese version of wiktionary). Most of the errors are related to the mis-syllabification of glide plus vowel sequences (90.6%), which, according to the majority of the phonologists, belong to different syllables [1–3, 6]. We also found a relevant amount of entries which were empty in our retrieval (6.4%), i.e., they had some markdown formatting issue regarding their IPA transcription, as well as double spelling and different symbols representing the same sound (3%). All these errors where manually corrected and stored in a new file freely available on GitHub [1], which is intended to be merged into the current version of this resource.

Then, we performed initial statistics on the context-free frequency of the phonemes, which are presented in table 1. These results seem that the feature [coronal] is the most unmarked consonant feature in Portuguese [4], because even consonants like /t/ and /d/ which can only occupy one syllable position (the onset) are among the most frequent.

Table 1: Consonant frequency on Wikcionário

IPA	*count*	*%*
ɾ	8132	16,37
t	5546	11,16
d	3741	7,53
l	3580	7,21
s	3537	7,12
k	3525	7,1
p	2525	5,08
m	2500	5,03
ʃ	2321	4,67
n	1924	3,87
v	1486	2,99

IPA	*count*	*%*
b	1479	2,98
j	1364	2,75
w	1316	2,65
f	1315	2,65
z	1224	2,46
ʒ	1199	2,41
g	1193	2,4
ʀ	1121	2,26
ʎ	298	0,6
ɲ	208	0,42
k^w	140	0,28

Some advances have been made in order to search for distributional restrictions of the palatals /ʎ/ and /ɲ/, which display very low frequency [9]. However, future works aims at expanding the set consonants analysed and to make cross-linguitic comparison. For instance, the IPA transcriptions from Wikcionário were recently merged with PtLanka, a database of Sri Lanka Portuguese lexicon [8]. The resulting data set will enable the first large phonological comparison between a creole and its lexifier and, therefore, we expect it to be a major contribution to contact linguistics.

KEYWORDS

phonology, IPA transcription, Portuguese

ACM Reference Format:
Luís Trigo and Carlos Silva. 2022. Using wiktionary for advances on Portuguese phonology. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Acknowledgments

This research was conducted within the doctoral program of Languages Sciences (Faculty of Arts-University of Porto), was funded by the Portuguese Foundation for Science and Technology (FCT MCTES) through the PhD grant SFRH/BD/2020.07466.BD and supported by the Center of Linguistics of the University of Porto (FCT-UIDB/00022/2020).

REFERENCES

[1] Ernesto Andrade. 1998. Sobre a alternância vogal/glide em Português. In *Atas do XIII encontro nacional da Associação Portuguesa de Linguística*, Vol. 1. Lisboa, 91–102.

[2] Leda Bisol. 1999. A Sílaba e os seus constituintes. In *Gramática do Português Falado*, M. H. Neves (Ed.). Sao Paulo.

[*] Both authors contributed equally to this research.
[1] https://github.com/Portophon/wiktionary

[3] Lurdes Ferreira. 2014. Ditongos crescentes: um conceito fonológico ou fonético *Letras de Hoje* 49 (2014), 28–35. https://doi.org/10.15448/1984-7726.2014.1.14646

[4] Maria Helena Mateus and Ernesto Andrade. 2000. *The Phonology of Portuguese.* Oxford University Press.

[5] Christian M Meyer and Iryna Gurevych. 2012. *Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography.* na.

[6] Mariana Ribeiro and Carlos Silva. 2021. How truthful are Portuguese false diphthongs: an empirical approach. In *Vienna workshop on Portuguese Linguistics.*

[7] Cristina Romani, Claudia Galuzzi, Cecilia Guariglia, and Jeremy Goslin. 2017. Comparing phoneme frequency, age of acquisition, and loss in aphasia: Implications for phonological universals. *Cognitive Neuropsychology* 34 (2017), 449 – 471.

[8] Luis Trigo and Carlos Silva. 2021. PtLanka: an online corpus of Sri Lanka Portuguese lexicon and phonology. In *OpenCor 2021.*

[9] Luís Trigo and Carlos Silva. forc. Comparing lexical and usage frequencies of palatal segments in Portuguese. In *Proceedings of the 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022).* Springer, Fortaleza.