Using a hybrid approach to build a pronunciation dictionary for Brazilian Portuguese

Gustavo Mendonça, Sandra Aluisio

Instituto de Ciências Matemáticas e de Computação
University of São Paulo, Brazil
gustavom@icmc.usp.br, sandra@icmc.usp.br

10.21437/Interspeech.2014-319

Abstract

This paper describes the method employed to build a machine-readable pronunciation dictionary for Brazilian Portuguese. The dictionary makes use of a hybrid approach for converting graphemes into phonemes, based on both manual transcription rules and machine learning algorithms. It makes use of a word list compiled from the Portuguese Wikipedia dump. Wikipedia articles were transformed into plain text, tokenized and word types were extracted. A language identification tool was developed to detect loanwords among data. Words' syllable boundaries and stress were identified. The transcription task was carried out in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict the transcription of the remaining graphemes (mostly vowels). The method was evaluated through 5-fold cross-validation; results show a F1-score of 0.98. The dictionary and all the resources used to build it were made publicly available.

Index Terms: pronunciation dictionary, grapheme to phoneme conversion, text to speech

1. Introduction

In many day-to-day situations, people can now interact with machines and computers through the most natural human way of communication: speech. Speech Technologies are present in GPS navigation devices, dictation systems in text editors, voice-guided browsers for the vision-impaired, mobile phones and many other applications [1]. However, for many languages, there is a dire shortage of resources for building speech technology systems. Brazilian Portuguese can be considered one of these languages. Despite being 6th most spoken language in the world [2], with about 200 million speakers, speech recognition and speech synthesis for Brazilian Portuguese are far from the current state of the art [3]. In this paper, we describe the method employed in building a publicly available pronunciation dictionary for Brazilian Portuguese which tries to diminish this scarcity.

The dictionary makes use of a hybrid approach for grapheme to phoneme conversion, based on both manual transcription rules and machine learning algorithms, and aims at promoting the development of novel speech technologies for Brazilian Portuguese. Hybrid approaches in grapheme to phoneme conversion have been applied successfully to other languages [4][5][6][7]. They have the benefit of taking advantage from both knowledge-based and data-driven methods. We propose a method in which the phonetic transcription of a given word is obtained through a two-step procedure. Its primary word list derives from the Portuguese Wikipedia dump of 23rd January 2014. We decided to use Wikipedia as the primary word list for the dictionary for many reasons: i) given its encyclopedia nature, it covers wide-ranging topics, providing words from both general knowledge and specialized jargon; ii) it contains around 168,8 million word tokens, being robust enough for the task; iii) it makes uses of crowdsourcing, lessening author's bias; iv) its articles are distributed through Creative Commons License. Wikipedia articles were transformed into plain text, tokenized and word types were extracted.

We developed a language identifier in order to detect loanwords among data. It is a known fact that when languages interact, linguistic exchanges inevitably occur. One particular type of linguistic exchange is of great concern while building a pronunciation dictionary, namely, non-assimilated loanwords [8]. Non-assimilated loanwords stand for lexical borrowings in which the borrowed word is incorporated from one language into another straightforwardly, without any translation or orthographic adaptation. These words represent a problem to grapheme-to-phoneme (G2P) conversion since they show orthographic patterns which are not predicted in advance by rules or which are too deviant to be captured by machine learning algorithms. Many algorithms have been proposed to address Language Identification (LID) from text [9][10][11][12]. Since our goal is to detect the language of single words, we employed n-gram character models in the identifier, given its previous success in dealing with short sequences of characters.

Brazilian Portuguese Phonology can be regarded as syllable and stress-driven [13]. In fact, many phonological processes in Brazilian Portuguese are related to or conditioned by syllable structure and stress position [14]. Vowel harmony occurs in pretonic context [15], posttonic syllables show a limited vowel inventory [13], nasalization occurs when stress syllables are followed by nasal consonants [16], epenthesis' processes are triggered by the occurrence of non-allowed consonants in coda position [17] and so on and so forth. Therefore, detecting syllable boundaries and stress is of crucial importance for G2P systems, in order to achieve correct transcriptions. Several algorithms have been proposed to deal with the syllabification in Brazilian Portuguese. However most of them were not extensively evaluated nor were made publicly available [18] [19] [3] [20]. For this reason, we implemented our own syllabification algorithm, based directly on the rules of the last Portuguese Language Orthographic Agreement [21].

Word types recognized as belonging to Brazilian Portuguese by the language identifier were transcribed in a two-step process: i) words are submitted to a set of transcription rules, in which predictable graphemes (mostly consonants) are transcribed; ii) a machine learning classifier is used to predict

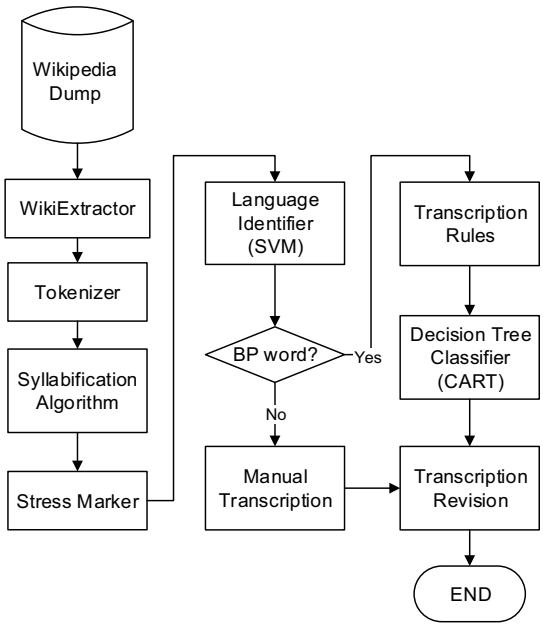

Figure 1: *System architecture for building the pronunciation dictionary.*

the transcription of the remaining graphemes (mostly vowels). All the data were subsequently revised. Figure 1 summarizes the method.

2. Method

2.1. Primary Word List

We used the Portuguese Wikipedia's dump of 23rd January 2014 as the primary word list for the pronunciation dictionary. In order to obtain plain text from the articles, we employed WikiExtractor [22]; it strips all the MediaWiki markups and metadata forms. Afterwards, texts were tokenized and unique words types extracted. The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820,000 articles. With the purpose of avoiding misspellings, URLs and other spurious data, only words with frequency higher than 10, which showed neither digits nor punctuation marks were selected.

2.2. Language Identifier

A Language Identifier module was developed in order to detect loanwords in the pronunciation dictionary. The Identifier consists of a Linear Support Vector Machine Classifier [23] and was implemented in Python, through Scikit-learn [24]. It was trained on a corpus made of the 200,000, containing 100,000 Brazilian Portuguese words and 20,000 words of each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. We selected these languages because they are the major donors of loanwords to Brazilian Portuguese [25]. From these words we extracted features such as initial and final bi- and trigraphs; number of accented graphs, vowel-consonant ratio; average mono-, bi- and trigraphs probability; and used them to estimate the classifier. Further details can be found in the website of the Project[1]. After training, we applied the classifier to the Wikipedia word list with the purpose of identifying loanwords among data. The identified loanwords were then separated from the rest of words for later revision, i.e. they were not submitted to automatic transcription.

2.3. Syllabification algorithm and stress marker

Our syllabification algorithm follows a rule-approach and is based straightforwardly on the syllabification rules described in the Portuguese Language Orthographic Agreement [21]. Given space limitations, rules were omitted from this paper as they can be found in the website of the project, along with all the resources developed for the dictionary. As for the stress marker, once the syllable structure is known in Brazilian Portuguese, one can predict where stress falls. Stress falls:

1. on the antepenultimate syllable if it has an accented vowel <á,â,é,ê,í,ó,ô,ú>;
2. on the ultimate syllable if it contains the accented vowels <á,é,ó> or <i,u>; or if it ends with one of the following consonants <r,x,n,l,z>;
3. on the penultimate syllable otherwise.

2.4. Transcriber

The transcriber is based on a hybrid approach, making use of manual transcription rules and an automatic classifier, which builds Decision Trees. Initially, transcription rules are applied to the words. The rules covers not all possible graphemes to phoneme relations, but only those which are predictable by context. The output of the rules is what we called the intermediary transcription form. After obtaining it, a machine learning classifier is applied in order to predict the transcription of the remaining graphemes. Figure 2 gives an example of the transcription process.

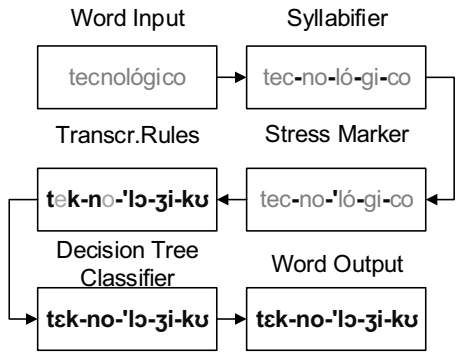

Figure 2: *Example of the transcription procedure – in grey: graphemes yet to be transcribed; in black: graphemes already transcribed.*

The rules' phase has two main goals: guarantee the correct transcription of certain predictable graphemes (mostly consonants) and also ensure the alignment between graphemes and phones for the classifier. They were set in order to avoid overlapping and order conflicts. Long sequences of graphemes, such

[1] http://nilc.icmc.usp.br/listener/aeiouado

as triphthongs, contextual diphthongs and general diphthongs are transcribed first (e.g. <x-ce>→[-se]). Then graphemes involving phones that undergo phonological processes are transcribed (e.g. <ti>→[tʃi], <di>→[dʒi]). After that, several contextual and general monophones are transcribed (e.g. <#x>→[ʃ], <#e-x>→[#e-z]).

On what regards to the classifier, it was developed primarily to deal with the transcription of vowels. In Brazilian Portuguese, vowels have a very irregular behavior, specially the mid ones. Therefore the relations between the vowels' graphemes and their corresponding phonemes are hard to predict beforehand through rules. Consider, for instance, the words "teto" *(roof)* and "gueto" *(ghetto)*; both are nouns and share basically the same orthographic environment. However the former is pronounced with an open "e" ['tɛ.tʊ] and the latter with a closed one ['ge.tʊ]. The classifier employs Decision Trees, through an optimised version of the CART (Classification and Regression Trees) algorithm and was implemented in Python, by means of the Scikit-learn library [24].

The algorithm was trained over a corpus of 3,500 words phonetically transcribed and manually revised, with a total of 39,934 instances of phones. The feature extraction happened in the following way. After reviewing the data, we obtained the intermediary transcription form for each of these words and aligned them with the manual transcription. Then, we split the intermediary transcription form into its corresponding phones and, for each phone, we extracted the following information: i) the phone itself; ii) 8 previous phones; iii) 8 following phones; iv) the distance between the phone and the tonic syllable; v) word class – parts of speech; v) the manually transcribed phone. We considered a window of 8 phones in order deal with vowel harmony phenomena. By establishing a window with such length, one can assure that pretonic phones will be able to reach the transcription of the vowels in the stressed syllable. The classifier was applied to all 108,389 words categorized as BP words by the Language Identifier module, all of them were cross-checked by two linguists with experience in Phonetics and Phonology.

3. Results

The Portuguese Wikipedia has about 168,8 million word tokens and 9,7 million types, distributed among 820k articles. After applying the filters to the data, i.e. words with frequency higher than 10, with no digits nor punctuation marks, we ended up with circa 238k word types, representing 151,9 million tokens. Table 1 describes the data.

Table 1: *Portuguese Wikipedia Summary – Dumped on 23rd January 2014.*

	Word Tokens	**Word Types**
Wikipedia	168,823,100	9,688,039
Selected	151,911,350	238,012
% Used	90.0	2.4

The selected words covers 90,0% of the Wikipedia content. Although the number of selected word types seems too small at first glance, one of the reasons is that 7,901,277 of the discarded words were numbers (81,5%). The remaining discarded words contained misspellings (*dirijem-se* – it should be *dirigem-se*), used a non-Roman alphabet (*λόγω*), were proper names (*Stolichno*, *Zé-pereira*), scientific names (*Aegyptophitecus*), abbreviations or acronyms (LCD, HDMI).

As for the language identifier, we trained and evaluated it with the 200,000 words multilingual corpus. The corpus consists of 100,000 Brazilian Portuguese words and 20,000 words from each of the following languages: English, French, German, Italian and Spanish. All of these words were collected through web crawling News' sites and were not revised. The results obtained for the identifier, through 5-fold cross validation are described in Table 2.

Table 2: *Results from the Language Identifier module – Training Phase.*

	Precision	**Recall**	**F1-score**	**Support**
BP words	0.85	0.89	0.87	100,000
Foreign Words	0.88	0.84	0.86	100,000
Avg/Total	0.86	0.86	0.86	200,000

The classifier showed an average F1-score of 0.86. Although such result is not as good as we expected – some authors reported 99% by using similar methods with trigrams probability, the relatively low F1-score can be explained given the nature of the data. In most language identifiers, the input consists of texts or several sentences, in other words, there is much more data available for the classifier. Since we are working with single words, the confusion of the model is higher and the results are, consequently, worse. Additionally, because the word list used to train the identifier was not revised, there is noise among the data. After training and evaluating the classifier, we applied it to the selected word list derived from the Wikipedia, in order to detect loanwords. Table 3 describes the results gathered.

Table 3: *Results from the Language Identifier module – Wikipedia word list.*

	Wikipedia word list
BP words	108,370 (46%)
Foreign Words	129,642 (54%)
Total	238,012

As one can observe, although we established a frequency filter to avoid spurious words, many loanwords still remain. More than half of the word list selected from Wikipedia consists of foreign words. Notwithstanding that, the list of Brazilian Portuguese words is still of considerable size. For instance, the CMUdict [26], a reference pronunciation dictionary for the English language, has about 125,000 word types.

Concerning the syllabification algorithm and the stress marker, we did not evaluate them in isolation, but together with the transcriber since the rules for each of these modules are intertwined. That is to say the transcription rules are strictly dependent on the stress marker module and the syllable identifier. Besides, the Decision Tree Classifier is built upon the output of the transcription rules, so it is entirely dependent on it. The Decision Tree Classifier was trained over a corpus of 3,500 cross-checked transcribed words, containing 39,934 instances of phones. We analyzed its performance through 5-fold cross validation, the results for each individual phone are summarized in Table 4.

As it can be seen, the method achieved very good results, with a F1-score of 0.98. Many segments were transcribed with 100% accuracy, most of them were consonants. As it was expected, the worst results are related to mid vowels [ɛ, e, ɔ, o],

Table 4: *Results from the Transcriber – Training Phase.*

	Precision	Recall	F1-score	Support
syl. boundary	1.00	1.00	1.00	9099
stress	1.00	1.00	1.00	3507
p	1.00	1.00	1.00	760
b	1.00	1.00	1.00	357
t	0.99	0.99	0.99	1135
d	0.99	0.99	0.99	1148
k	0.99	0.99	0.99	978
g	1.00	1.00	1.00	298
tʃ	0.98	0.98	0.97	450
dʒ	0.96	0.96	0.96	243
m	1.00	1.00	1.00	668
n	1.00	1.00	1.00	556
ɲ	1.00	1.00	1.00	69
f	1.00	1.00	1.00	311
v	1.00	1.00	1.00	531
s	0.98	0.98	0.98	2309
z	0.93	0.94	0.93	416
ʃ	0.84	0.84	0.84	138
k.s	0.72	0.64	0.66	41
ʒ	1.00	1.00	1.00	196
l	1.00	1.00	1.00	682
ʎ	1.00	1.00	1.00	58
ɾ	1.00	1.00	1.00	1388
h	0.98	0.99	0.99	737
fi	0.97	0.92	0.94	169
w	0.97	0.98	0.97	441
w̃	0.98	0.99	0.99	309
j	0.97	0.95	0.96	223
j̃	0.95	1.00	0.98	110
a	1.00	1.00	0.99	2316
ə	0.99	0.99	0.99	1093
ɛ	0.65	0.68	0.66	275
e	0.93	0.91	0.92	1779
i	0.98	0.99	0.98	2073
ɪ	0.97	0.97	0.97	365
ɔ	0.69	0.75	0.71	220
o	0.93	0.92	0.93	1112
u	0.96	0.96	0.96	488
ʊ	1.00	1.00	1.00	1033
ã	1.00	1.00	1.00	719
ẽ	0.96	0.97	0.97	497
ĩ	0.99	0.99	0.99	274
õ	0.97	0.96	0.97	299
ũ	0.94	0.92	0.93	64
Avg/Total	0.98	0.98	0.98	39934

specially mid-low vowels, [ɛ] showed a F1-score 0.66 and [ɔ] of 0.71. It can be the case that since the grapheme context is the same for [ɛ, e] and [ɔ, o], the Decision Tree classifier generalizes, in some cases, to the most frequent phone, that is the mid-high vowels [e,o]. The transcriber also had problems with the [k.s] (F1-score: 0.66) and [ʃ] (F1-score: 0.84). This result was also expected, both these phones are related to the grapheme <x> which, in Brazilian Portuguese, shows a very irregular behavior. In fact, <x> can be pronounced as [ʃ, s, z, k.s], depending on the word: "bruxa" *(witch)* [ʃ], "próximo" *(near)* [s]; "exame" *(test)* [z] and "axila" *(armpit)* [k.s].

4. Final Remarks

We presented the method we employed in building a pronunciation dictionary for Brazilian Portuguese. High F1-score values were achieved while transcribing most of the graphemes in Brazilian Portuguese and the dictionary can be considered robust enough for Large Vocabulary Continuous Speech Recognition (LVCSR) and Speech Synthesis. Although the rules we developed are language-specific, the architecture we used for compiling the dictionary, by using transcription rules and machine learning classifiers, can be successfully replicated in other languages. In addition, the entire dictionary, all scripts, algorithms and corpora were made publicly available.

5. Acknowledgements

Part of the results presented in this paper were obtained through research activity in the project titled "Semantic Processing of Brazilian Portuguese Texts", sponsored by *Samsung Eletrônica da Amazônia Ltda.* under the terms of Brazilian federal law number 8.248/91.

6. References

[1] R. Godwin-Jones, "Emerging technologies: Speech tools and technologies," *Language Learning and Technology*, vol. 13-3, pp. 4–11, 2009.

[2] F. Lewis, M. Gary and D. Charles, *Ethnologue: Languages of the World, Seventeenth edition*, ser. Seventeenth edition. Dallas, Texas: SIL International, 2013. [Online]. Available: http://www.ethnologue.com

[3] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[4] R. I. Damper, Y. Marchand, M. Adamson, and K. Gustafson, "Comparative evaluation of letter-to-sound conversion techniques for english text-to-speech synthesis," in *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis*, 1998.

[5] T. Polyakova and A. Bonafonte, "Learning from errors in grapheme-to-phoneme conversion." in *INTERSPEECH*, 2006.

[6] A. Teixeira, C. Oliveira, and L. Moutinho, "On the use of machine learning and syllable information in european portuguese grapheme-phone conversion," in *Computational Processing of the Portuguese Language*. Springer, 2006, pp. 212–215.

[7] A. Veiga, S. Candeias, and F. Perdigão, "Developing a hybrid grapheme to phoneme converter for european portuguese," vol. 1, pp. 297–300, May 2013.

[8] H. Bussmann, G. Trauth, K. Kazzazi, and H. Bussmann, *Routledge dictionary of language and linguistics / Hadumod Bussmann ; translated and edited by Gregory Trauth and Kerstin Kazzazi*. Routledge, London ; New York :, 1996.

[9] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific twitter collections," in *Proceedings of the Second Workshop on Language in Social Media*, ser. LSM '12. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012, pp. 65–74. [Online]. Available: http://dl.acm.org/citation.cfm?id=2390374.2390382

[10] E. B. Bilcu and J. Astola, "A hybrid neural network for language identification from text," in *Machine Learning for Signal Processing, 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on*. IEEE, 2006, pp. 253–258.

[11] D. Trieschnigg, D. Hiemstra, M. Theune, F. de Jong, and T. Meder, "An exploration of language identification techniques for the dutch folktale database," in *Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage, LREC 2012*, P. Osenova, S. Piperidis, M. Slavcheva, and C. Vertan, Eds. Istanbul, Turkey: LREC organization, May 2012, pp. 47–51. [Online]. Available: http://doc.utwente.nl/82013/

[12] M. Zampieri, B. G. Gebre, and H. Nijmegen, "Automatic identification of language varieties: The case of portuguese," in *Proceedings of KONVENS*, 2012, pp. 233–237.

[13] T. C. Silva, *Fonética e fonologia do português: roteiro de estudos e guia de exercícios*. Contexto, 2005.

[14] C. Girelli, *Brazilian Portuguese Syllable Structure*. UMI, 1990. [Online]. Available: http://books.google.com.br/books?id=KRGMnQEACAAJ

[15] L. Bisol, "Vowel harmony: a variable rule in brazilian portuguese," *Language Variation and change*, vol. 1, pp. 185–198, 1989.

[16] A. Quicoli, "Harmony, lowering and nasalization in brazilian portuguese," *Lngua*, vol. 80, pp. 295–331, 1990.

[17] F. Delatorre and R. Koerich, "Production of epenthesis in ed-endings by brazilian efl learners," *Proceedings of the II Academic Forum*, p. 8, 2005.

[18] C. Oliveira, L. C. Moutinho, and A. J. S. Teixeira, "On european portuguese automatic syllabification." in *Proceedings of the Interspeech 2005*. ISCA, 2005, pp. 2933–2936. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2005/i05_2933.html

[19] V. Vasileʋski, "Phonologic patterns of brazilian portuguese: a grapheme to phoneme converter based study," in *Proceedings of the Workshop on Computational Models of Language Acquisition and Loss*. Avignon, France: Association for Computational Linguistics, April 2012, pp. 51–60. [Online]. Available: http://www.aclweb.org/anthology/W12-0912

[20] W. Rocha and N. Neto, "Implementação de um separador silábico gratuito baseado em regras linguísticas para o português brasileiro," in *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 2013, pp. 108–115.

[21] Brasil, *Acordo ortogrfico da lngua portuguesa, de 14, 15 e 16 de dezembro de 1990*. Braslia, Brazil: Dirio do Congresso Nacional da Repblica Federativa do Brasil, Poder Executivo, 2009.

[22] Medialab, "Wikipedia extractor," http://medialab.di.unipi.it/wiki/Wikipedia_Extractor, 2013.

[23] I. Steinwart and A. Christmann, *Support vector machines*. Springer, 2008.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] I. Alves, *Neologismo: Criação lexical*, ser. Princípios (São Paulo). Editora Atica, 2001. [Online]. Available: http://books.google.com.br/books?id=7fIuAAAAYAAJ

[26] H. Weide, "The cmu pronouncing dictionary," 1998. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict