PETRUS: A rule-based grapheme-to-phone converter for Brazilian Portuguese

Vanessa Marquiafável\(^1\), Alessandro Bokan\(^2\), Claudia Zavaglia\(^3\)

\(^1\) Universidade Estadual Paulista (IBILCE-UNESP)
\(^2\) Universidade de São Paulo (ICMC-USP)
\(^3\) Núcleo Interinstitucional de Linguística Computacional (NILC)

[marquiafavel,alessandro.bokan]@gmail.com
zavaglia@ibilce.unesp.br

Abstract. This paper describes a grapheme-to-phone system. One of the major challenges for these types of systems is to give the correct phonetic symbol from a sequence of characters. The justification for developing a new G2P for Brazilian Portuguese arises from factors such as the little sharing of systems already implemented algorithms, and the results of performed tests from the studies that points out conversion errors that persist in currently available systems. In this context, it was developed a grapheme-to-phone system named PETRUS, which provides automatic phonetic transcription through a web interface in order to support automatic phonetic transcription of lemmas for Brazilian dictionaries’ entries.

1 Introduction

This paper aims to briefly describe a web based grapheme-to-phone (G2P) system for Brazilian Portuguese (BP). Studies in speech processing show much promise and the trend of man-machine interaction seems to be leaning evermore towards speech commands, rather than mice and keyboards; so that, we expect they offer satisfactory performances. One of the modules from a system like that is the G2P. It is responsible for the process of automatically converting a string of graphemes (orthographic character) into a sequence of phonetic (or phonological) symbols. Previous work on G2P can be pointed out for European Portuguese (EP) \([1, 2, 3]\) and also for BP \([4, 5, 6]\). The justification for developing a new one for BP arises from two main factors, such as the little sharing of systems already implemented algorithms and the results of performed tests from the studies that points out conversion errors that persist in currently available systems. Besides this, there is not any G2P developed until now that considers São Paulo standard dialect, the most widespread one.

In this context, we present here PETRUS, a web-based G2P system to automatically convert a sequence of letters such as "<descrédito>" into a sequence of IPA (International Phonetic Alphabet) symbols, \([dʒiskrɛʤiðu]\). It is a computational environment to support automatic phonetic transcription of lemmas (isolated lexical units) for Bra-
zilian dictionaries' entries. Our main objective is to give support for lexicographers in adding phonetic information in most Brazilian dictionaries that they produce. The dictionaries intended for native speakers do not usually include phonetic transcription. Just some orthoepic information would appear close to the headwords when it is absolutely necessary. The problem intensifies when the dictionary user is not a native speaker of Portuguese. Unlike the other existing G2P, PETRUS uses information from syllabification, stressed vowel and grammatical classes to perform word’s phonetic transcription in order to improve hit rate. BP comprises several regional dialects, which can be roughly identified with the boundaries of the States of the federation. It is further possible to subcategorize these dialects into urban and rural. The most important television news announcers are trained to use what is referred to as less socially marked pronunciation. In practice, they use a modern variety of the Paulista dialect, called Paulista standard [7, 8]. As in all varieties of a language, this dialect presents different pronunciations for particular contexts or even for a single word. This requires a great deal of conventions, which should not betray a good pronunciation, but on the other hand, should show some generalization, convenient for a BP G2P system.

2 Overview of PETRUS system

PETRUS is composed by six modules. The first module is responsible for phonetic transcribing homographic heterophone words (similar in writing, but different in speaking, i.e., ‘g[o]sto’ (noun) e ‘g[O]sto’ (verb)) and the linguistic resource used to built this module was a dictionary with almost 1.000 homograph heterophone pairs transcribed. The second one, for identifying the presence of prefix in a word before transcribing it, because prefixes in BP have a specific behavior towards vowel alternation, and the linguistic resource used was a rule-based algorithm with 145 rules. Then, the next module identifies with an ’[’ the stressed vowel of the word that is being analyzed, using a rule-based algorithm from [9]. Fourth module identifies and marks syllable boundaries and use 25 rule-based algorithms from [9]. Fifth module performs a Part-of-Speech tagging using UNITEX [10] and POS tagger from [11]. Only then, on the last module, all this linguistic information is used in order to phonetically transcribe a word, by the use of a set containing 160 rule-based algorithms (62 to transcribe consonants and 98 rules for vowels). The main advantages of using a rule-based approach is that the AG2P is always able to transcribe a new word and that any approach that has used statistical methods or machine learning performed better than 98% of correctly transcribed phones. Petrus can be accessed at http://nilc.icmc.usp.br/nlcn/index.php/tools-and-resources.

3 Conclusions and future work

The main goal of PETRUS system is to phonetically transcribe words according to Paulista standard dialect. The system described here is the first release of a web interface to access PETRUS online. Until this phase, regarding the analysis of quality in the phonetic transcriptions performed by it, we have conducted a small experiment in
which the phonetic transcription was evaluated with reference to a manual annotation corpus. It was not possible to compare it with the other G2P system for BP because they do not adopt the same dialect and are not all available for public use. In a near future, we will include the mechanisms to upload also a .txt file. This functionality will permit to phonetically transcribe a bigger input. Concerning the usability of the interface, we will also add alternative forms for users who want to use just the syllabic module, the tonic vowel or the POS tagger module.

4 Acknowledgments

This research has been funded by FAPESP (Process: 2011/01299-4).

5 References