# Morphosyntactic Parser for Brazilian Portuguese: Methodology for Development and Assessment

Izabel Christine Seara, Fernando Santana Pacheco,
Sandra Ghizoni Kafka, Rui Seara Jr., and Rui Seara*

LINSE – Circuits and Signal Processing Laboratory
Department of Electrical Engineering
Federal University of Santa Catarina, Brazil
{izabels, fernando, kafka, ruijr, seara}@linse.ufsc.br

**Abstract –** In text-to-speech (TTS) systems, an effective morphosyntactic classification is important to improve the prosody of synthesized speech as well as the pronunciation of words subject to vocalic alternation. This research work presents a methodology used for developing and assessing an ad hoc morphosyntactic parser to a TTS system for Brazilian Portuguese. The developed parser is composed of a dictionary and a set of rules structured in four levels. The methodology used for development consisted firstly in the creation of a large annotated dataset and an incremental development of rules for morphosyntactic classification. By using this approach, the achieved accuracy rate of the classification process is of 98.59% for words and 80.66% for sentences in a specific dataset.

**Keywords:** Brazilian Portuguese, morphosyntactic parser, text-to-speech.

## 1 Introduction

Currently, a number of automatic morphosyntatic parsers is conceived for different purposes, besides its application in text-to-speech (TTS) systems [1, 2, 3]. These systems (mainly those using concatenative speech synthesis) search for a better prosody through a more detailed linguistic description, avoiding artificially changing the acoustic parameters of the synthesized speech. Thereby, the evaluation of a parser developed with such a goal would not be easily applied to automatic parsers that present a more wide-ranging linguistic description, since the labels differ from each other and the details required by the parsers with nonrestrictive aims are distinct from the ones which are needed for the present application. However, such detail levels also could be included in our parser without any difficulties, once that information is incorporated in the TTS dictionary.

*This work was partially supported by the Brazilian National Council for Scientific and Technological Development (CNPq), the Committee for Postgraduate Courses in Higher Education (CAPES), and Dígitro Tecnologia Ltda.

Now focusing on TTS systems, we have verified that the quality of these systems has improved considerably over the past few years, making possible the use of speech synthesis for a large number of applications. TTS systems are used as, for example, to generate speech output in spoken dialog systems and call centers. Although recent advances have enhanced some synthetic speech characteristics, a few drawbacks still impact negatively the overall TTS quality. Here, an overview of the processing chain of TTS systems is important to better understand the source of such problems.

A general TTS system has two basic processing stages: the linguistic one and the speech signal synthesis *per se*. The system considered here has four sub-modules related to linguistic features. The first is a grapheme to phoneme (G2P) converter, which maps the letters in the text to phonetic symbols. The second module is responsible for the syllabic division, which is a linguistic parameter significant for the prosodic aspect. In the third module, a morphosyntactic parser labels each lexical item into a morphological class. Aiming to improve the naturalness, in the fourth module, pauses are inserted based on information from the parser.

The morphosyntactic classification is a fundamental component of TTS systems. In the open literature, different strategies are discussed, from purely statistic-based methods to hand-made rule based systems [4, 5]. Besides other uses, the morphosyntactic classification is needed for assigning the correct phoneme in homograph words, which can present vocalic alternation [6]. This characteristic is related to the open or closed forms for the vowels *e* and *o* in some Brazilian Portuguese words. Some homograph non-homophone words can be distinguished only with the knowledge of the syntactic classification. For instance, note that in the sentence *eu olho para o problema com o meu olho clínico* (I look at the trouble with my clinical eye) the first occurrence of *olho* (['ɔʎʊ]) (I look at) is a verbal form of the verb *olhar* (to look at) and it is spoken with the open vowel [ɔ]. In the second, *olho* (['oʎʊ]) (eye) is a noun and it is spoken with the close vowel [o]. An error of this type may reduce the global perceived quality of the TTS system. G2P rules usually assign the correct phoneme for the most part of the words [7]. However, we have noted that the ambiguity in homograph non-homophone words can be solved only with the classification of the full sentence. A partial classification is insufficient, since errors in the classification of some lexical items can lead to errors in crucial words (in this case, the homograph non-homophone words). Thus, the main question to solve would be "how to develop a robust parser capable of classifying correctly all words of a sentence?"

In this research work, our main concern is with the parser assessment, since the evaluation that we had so far was only a subjective one. Thereby, our goal is to present a methodology for objective assessment of a morphosyntactic parser for a TTS system. Such an approach has improved our confidence in the overall quality of the morphosyntactic parser.

This paper is organized as follows. An overview of the morphosyntactic parser and some classification problems are discussed in Section 2. Methodology for assessment of the used rules is described in Section 3. Section 4 presents the assessment results of the morphosyntactic parser. Remarks and some suggestions for future work are given in Section 5.

## 2 Morphosyntactic Parser

Different strategies are discussed in the literature for the implementation of a morphosyntactic parser [1, 2, 3, 8, 9]. We present here the reasoning taken into account to select our approach. Statistic-based parsers usually require a large and correctly labeled corpus. As another option would be a parser based on generative grammar but this could not be the best alternative, since a large labeled corpus with morphological details such as verbal tense or singular/plural nouns is also required. In this way, considering the unavailability of a large annotated corpus during development of the first version of our TTS system, we opt to develop a rule-based parser[1].

Two main components form it: a dictionary and rules. In the dictionary, we have included:

- closed-class words (those without ambiguity w.r.t. the morphosyntactic classification);
- flexioned verbal forms, important to solve cases of vocalic alternation;
- foreign words;
- words with more than one morphosyntactic classification (treated with a specific procedure).

Currently, this dictionary stores more than 60,000 entries. Fifty per cent of them have a full morphological classification.

From the initial classification given by the dictionary, rules analyze the context of each lexical item and assign a class to it. Therefore, the final result is a complete classification of the elements of a sentence w.r.t. the syntactic class as well as the morphology (gender and number, for nouns; person and tense, for verbs; among others). Before discussing the rules, we give an overview of the classes in the following section.

### 2.1 Morphosyntactic Classes

The following word classes are defined here for Brazilian Portuguese [8]:

**Table 1.** Morphosyntactic parser labels with some examples

| Class | Label | Example |
|---|---|---|
| *Noun* | [NOME] | João, quadro |
| *Adjective* | [ADJ] | bonito |
| *Verb* | [VER]/ [VERN] | é, deu/ dar, dado, dando |
| *Pronoun* | [RET]/ [OBA]/ [OBT]/ [POS]/ [DEM]/ [REL]/ [IND]/ [DEM1]/ [IND1]/ [LREL] | eu/o/lhe/meu/este/que/todos/ isto/tudo/a qual |
| *Article* | [ART] | as, um |
| *Preposition* | [PREP]/ [LPREP] | de/através de |
| *Conjunction* | [CONJ]/ [LCONJ] | quando/até que |
| *Adverb* | [ADV]/ [LADV] | hoje, no momento |
| *Numeral* | [NUM] | dois, terceiro |
| *Other classes* | [PER]/ [PDEN]/ [INT] | Quem?/Que lindo!/Oba! |

[1] Nowadays, with the availability of a corpus such as Floresta Sintá(c)tica (for Brazilian and European Portuguese) [3], a statistical method could be considered.

### 2.2 Levels of Rules in the Parser

The rules for morphosyntactic classification are based on the context surrounding each lexical item. Thus, a word is classified considering the two or three items before or after it (see examples in Section 3). To classify the words, the morphosyntactic parser in the TTS system is structured as follows. Firstly, items that appear in the dictionary with only one class are labeled. After that, a module with four levels analyzes the words. In the first level, 12 rules are used to pre-label lexical items that present verbal root and verbal ending as [VER] or [VERN]. In the second level, formed by 253 rules, lexical items corresponding to locutions ([LREL], [LADV], [LCONJ] [2], or [LPREP][3]) are labeled. Since some of the words classified in the first level can work either as VER/VERN or ADJ/NOME, the third level copes with such ambiguities. A total of 370 rules is applied in this level, considering the position of the words in the sentence. Finally, the fourth level (with 14 rules) classifies sentences with a WH question. All lexical items that belong to closed classes (the ones that possess a limited number of items, namely articles, pronouns, prepositions, conjunctions, and more often adverbs) have been inserted into the dictionary. When a lexical item belongs to more than one class, its ambiguity must be removed. This procedure is carried out through rules that consider the position and the interaction of the item with other words in the sentence.

In the verb classification, verbal roots and verbal endings have been inserted in a dictionary. When the word to be classified has verbal root and/or verbal ending, it is labeled as [VER] or [VERN], depending on the verbal ending. If such an item belongs to more than one class, it is inserted in a dictionary termed [NOMEVER]. The items that compose this dictionary will be reclassified in a higher-level analysis through rules that assess its function in the sentence to be synthesized. For example, the word *modelo* (to model) must be classified as [VER] in the sentence *modelo a argila muito bem* [I model the clay very well] and as [NOME] in the sentence *minha irmã virou modelo* [my sister became a model].

## 3 Methodology for Rule Assessment

During our first attempt to implement the morphosyntactic parser, we note that it was impossible for a human expert to remember all considered rules. Moreover, with a larger number of rules, the inclusion of a new one can create some kind of side effect. Whereas one particular problem in a sentence has been solved, some unexpected errors can also arise. To solve such problems, another methodology should be applied.

Thereby, the adopted strategy consists in using a medium-size dataset[4], manually annotated by an expert, and creating classification rules from this dataset. This corpus is composed of 13,330 sentences with 205,813 words. For training, this corpus has been accurately hand tagged with 25 tags (see Table 1). After training over 85% of

[2] LREL (for example: as quais, aos quais); LADV (for example: no meio de); LCONJ (for example: à medida que).

[3] LPREP is resulting of contractions of PREP+ ART, as for example: do (de+o) or no (em+o).

[4] When compared with corpora available in Linguateca (http://www.linguateca.pt/).

that corpus (11,320 sentences), we use the remainder, i.e., 2,010 sentences, for the test procedure (assessment phase). This enhancement approach includes three stages. In the first, statistical data is collected from the manually annotated dataset. The most frequent class for a given lexical item is considered as default for items with more than one possible class. Thus, even situations not covered by the rules can receive a classification at the end of the process. In the second stage, a histogram of incorrect classifications, obtained by comparing parser results with those from the manually classified dataset, is assessed. Finally, these statistics are considered in the third stage, in which rules are created and refined taking into account the most frequent contexts. For each new inserted rule, the process is repeated, incrementally improving the classification accuracy. Currently, we have a total number of 649 rules distributed within the four-level parser. Now, we show an illustrative example of ambiguity problems. For example, the lexical item *que* can be labeled as [REL] or [CONJ]. Additional results are presented in Section 4.

(1) *A menina que (REL) foi à festa resolveu o enigma*.
(The girl that went to the party has solved the enigma.)
(2) *Ele disse que (CONJ) chegaria tarde.*
(He said that he would arrive late.)

## 4 Parser Evaluation

Before applying the approach described in Section 3, the accuracy rate obtained by the parser was 90.76% (words) and 30.46% (sentences). After the refining process the accuracy rate becomes 98.59% (words) and 80.66% (sentences). In addition, to confirm the better classification performance, we assess the parser by using the test dataset for evaluation (2,010 sentences). The accuracy rate obtained is 82.03% for sentences and 99% for words, showing that the increased performance is consistent.

## 5 Final Remarks

This work presented a methodology used for development and objective assessment of a morphosyntactic parser for Brazilian Portuguese. This parser is composed of a dictionary and a set of rules structured in four levels. The methodology used for development consisted firstly in the creation of a large annotated dataset and an incremental development of rules for classifying morphosyntactic classes. Results showed that the achieved accuracy rate in the classification process is of 98.59% for words and 80.66% for sentences, for a given dataset. For future work, we will use the collected statistics for refining the classification, creating a hybrid approach, rule and statistic-based.

## References

1. Bick, E.: A Constraint Grammar-Based Parser for Spanish. In: Proceedings of TIL 2006 – 4th Workshop on Information and Human Language Technology. Ribeirão Preto, October (2006)
2. Ribeiro, R. D., Oliveira, L. C., Trancoso, I.: Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. In PROPOR'2003 – 6th Workshop on Computational Processing of the Portuguese Language, pp. 143-150. Springer-Verlag, Heidelberg, Faro, Portugal, June (2003)
3. Afonso, S.: Árvores deitadas: Descrição do formato e das opções de análise na Floresta Sintáctica. Documentação principal, em constante atualização. Disponível em http://acdc.linguateca.pt/treebank/DocumentacaoFloresta.html#Bick 2000
4. Taylor, P.: Text-to-speech synthesis. Cambridge, Cambridge University Press (2009)
5. Tatham, M., Morton, K.: Developments in Speech Synthesis. Wiley (2005)
6. Seara, I. C., Kafka, S. G., Klein, S., Seara, R.: Vowel sound alternation of verbs and nouns of the Portuguese spoken in Brazil for application in text-to-speech synthesis. In: Journal of the Brazilian Telecommunications Society, vol. 17, no. 1, pp. 79-85. June (2002)
7. Silva, D. C., Lima, A. A., Maia, R., Braga, D., Moraes, J. F., Resende Jr. F. G. V.: A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing. In: Proc. of the International Telecommunications Symposium, pp. 992-996. Fortaleza, Brasil (2006)
8. Seara, I. C., Kafka, S. G., Seara Jr., R., Klein, S., Pacheco, F. S., Seara, R.: Pause insertion based on a morphosyntactic parser for Brazilian Portuguese text-to-speech systems. In: VI International Telecommunications Symposium (ITS2006), pp. 947-951 Fortaleza-CE, Brazil, September (2006)
9. Braga, D., Coelho, L., Resende Jr., F. G. V.: Homograph ambiguity resolution in front-end design for Portuguese TTS systems. In Interspeech 2007. Proceedings of the 8th Annual Conference of the InternationalSpeech Communication Association, pp. 1761-1764. Antwerp, Belgium, August (2007)