A Word Prosodic Algorithm for Brazilian Portuguese

Vera Vasilévski
Federal Universityof Technology – Paraná
Rua Cerejeiras, Bairro São Luís, ZIP 85892-000, Santa Helena, Paraná – Brazil

Leonor Scliar-Cabral
Federal Universityof Santa Catarina
Trindade, ZIP 88040-900, Florianópolis, Santa Catarina – Brazil

Márcio José Araújo
Federal Universityof Technology – Paraná
Rua Cerejeiras, Bairro São Luís, ZIP 85892-000, Santa Helena, Paraná – Brazil

Abstract

A presentation of a word rule-based algorithm that pinpoints the position of the word stressed syllable in a text, processes the data, and displays them as a statistical report. As a result, the output is the stress patterns of the respective language. This algorithm is one of a variety of tools available to researchers in Brazilian Portuguese. The algorithm focused on here works on the results of a grapheme to phoneme converter and syllable parser, which reads written texts and translates them into Brazilian Portuguese phonological patterns, by using the IPA symbols. This word algorithm may also be useful for research on Portuguese language stress, since, while decoding graphemes, it always preserves this particular prosody feature, that is, primary stress of words, to a great extent, dealing, in this case, with syllables, not phonemes. This feature analysis of two texts belonging to different genres illustrates the word algorithm working.

Keywords: Word prosodic algorithm, Brazilian Portuguese, Stress, Syllable.

1. Introduction

Writing is a human invention, a system to record permanently the oral messages. In this regard, some alphabetic systems, besides converting phonemes into graphemes, can also register word stress as used in Portuguese and Spanish systems, for example. In Portuguese, the main stress may fall on the last, penultimate or antepenultimate syllable. The Portuguese writing system records unambiguously those occurrences, thanks to the genius of the encoder Gonçalves Viana (1904) in the early twentieth century. He applied the principle of Occam's razor, and thus the pattern that has the highest frequency, i.e., the penultimate stress words that are spelled at the end of the word with "a", "o", "e", followed or not by "s", carry no graphic mark.

The invention of writing systems over 5,000 years ago showed little and slow improvement until the invention of phonographic systems, among them the alphabetic, the only one which has units, called graphemes (one or two letters in Brazilian Portuguese), which represent the phonemes, although the proto-alphabet system did not record all the word units, since it used an acronym consonantal writing. The Greeks improved the written representation of vowels (Cagliari, 2009), the syllabic summit which carries the most salient prosodic contrasts at the phonological word level.

Since Greek, as well as Latin, are languages in which prosodic contrasts at the phonological word level are marked by duration, such contrasts between long and short vowels are recovered by written morfo syntactic marks as the differences, in Latin, between the Nominative and Ablative endings in the first declension, or between the third and second thematic vowels of the respective conjugations, or by the rules of poetic metrics, like the difference between a dactyl (long, short, short), and a spondee (two long) (Scliar-Cabral, 2003). At least for the Indo-European languages, the most salient prosodic syllable would be marked somehow (Said Ali, 1964). Nevertheless, as was already mentioned, written systems are a very recent invention, characterized by their fixedness. The following comments will focus specifically on the alphabetic system.

Since the sociolinguistic varieties change rapidly and the written system should incorporate all the variants in space and time of the same phoneme of a given language, there is an ever increasing discrepancy between the graphemes and phonemes that they should represent. On the other hand, the letters of the alphabet are not enough to form the graphemes, which represent the phonemes, not to mention the fact that there is not a one-to-one relationship between them, by virtue of the distance between the oral and the written systems and also by the fact that criteria were used for the etymological spelling of some words. As a crucial example, we cite the case of vowels, just the phonemes that carry prosodic contrasts at the phonological word level: there are seven oral and five nasal vowels in Brazilian Portuguese, but only five letters to represent them (e, i, a, o, u), with the aggravating circumstance that among the last mentioned letters, some should also represent the glides /j/, /w/.
Diacritics, or graphic signals were used, such as acute symbol, and the caret to point out differences between vowels [+ low] and [-high, -low] (e.g. between é/ê, ó/ô, along with the tilde or digraphs consisting of a vowel letter followed by "m" or 'n" in the same syllable to represent nasal vowels. Many other digraphs are used for reasons of etymology, coinciding with a single as graphemes, representing the same phoneme like "ch"/"x", "ss"/"c" (before "e" , "i") and so on.

Despite the above mentioned irregularities for writing, for reading the alphabetic system of Brazilian Portuguese reaches around 95% transparency (Vasilévski, 2008, 2012), i.e., the predictability of reading in relation to spoken language is extremely high in this system. Moreover, the Portuguese system of signaling stress is even more transparent, since it reaches 99% of predictability (Vasilévski, 2008, 2012). Thus, once orthography is phonologically based, word rule-based algorithms should provide a good coverage of the association between graphemes and phonemes. However, probably no natural human writing system totally satisfies this assumption, because exceptions of conversion can be found in almost every language (Candeias, Perdigão & Veiga, 2011). From this concern, it seems plausible to consider that written data, once they reproduce those grapheme-phoneme correspondences and also stress, allow the study of various aspects of spoken language, among which are the ones related to prosody at the phonological word level, the focus of this work. Also, since the alphabetic principles are based on the phoneme representation, any automatic program must start from the phonological description of the respective language, which is the case of the present approach.

Thus, we describe here a word algorithm that identifies the position of the stressed syllable of the words in an input text, that processes this information and that features it as statistical report. The word algorithm addresses written language after its conversion into systematic oral language, i. e, converting graphemes into phonemes that cover all their possible variants in this language, using the symbols of the International Phonetic Alphabet (IPA, 2013), since phonological criteria predominate in alphabetic systems. In addition, the conversion covers Brazilian Portuguese spoken varieties, which differ to a great extent from European Portuguese varieties, as is well known. A sample of the word algorithm application, that is a word prosodic analysis of two Brazilian texts, belonging to distinct discourse genres, illustrates the procedures, as already mentioned. In this study, our goal is not to estimate any specific approach or model, nor do we use any phonological dictionary. Our work consists of translating written language rules into a computer language to see to what extent it is possible to reproduce the transparency attributed to Portuguese language system regarding translating graphemes into phonemes – signaling stressed vowels of words –, at first, and then regarding phonological and orthographic syllables – signaling stressed syllables of words – in this artificial environment. What we present here is a kind of refinement of the second part, that is, the information about the stressed syllables of texts treated and reported in numbers. For this reason, some aspects of our work are not developed in depth here, as they have been reported in previous studies (see Vasilévski, 2012a, 2011, 2010, Vasilévski, Scliar-Cabral & Araújo, 2013).

2. Word Prosody Of Brazilian Portuguese

Although nowadays some authors do not adopt some of the definitions and terminology that follow, they are classical and we consider them relevant for a better understanding of the tool we present here. Prosody – or the study of supra-segmental facts of language – comprises the phenomena of distinction that characterize and oppose units of speech longer than the isolated phoneme: syllable, metric foot, the phonological word, the clitic group, the phonological phrase, the into national phrase and the phonological utterance. Our algorithm deals with the phonological word, as already mentioned. Among its aims, Prosody investigates the pronunciation of words (Malmberg, 1993) (Phonetic Prosody) and it also refers to the ways to emphasize a word in an utterance, that is, the prominence. In Portuguese, the distribution of stress in a word allows the demarcation of its limits, while the melodic line allows the assignation of mode at the phonological utterance level (boundary tones) (Said Ali, 1964).

Stress results from a stronger expiration when pronouncing syllables, while the musical contrasts depend upon the relative intervals among frequencies produced by the vibrations of the vocal folds in the larynx, resulting in differences among questions, assertions, exclamations, etc. and also among emotional states (Said Ali, 1964). For example, in questions – which are marked in written Portuguese language by the interrogative signal "?" at the utterance ending, like in many other languages –, the intonation is always raised at the end of the utterance, if it is a yes/no question; if it is a wh question, the intonation is always falling. We address here the stress of the word that is understood as an emphasis given to a syllable, by means of the stress or pitch. Furthermore, such stress results from a greater expiratory strength or intensity in the emission of a vowel of a syllable, in contrast to the vowels of the other syllables (Câmara Jr, 1975).The phonological word in Portuguese is well defined, and its distinctive mark is stress (Câmara Jr, 1986). The phonological concept distinguishes stressed and non-stressed words. The stressed words are phonological words, and the non-stressed ones are the clitics (Câmara Jr, 1975). Brazilian Portuguese vowels are fully characterized by stress position and the phonological word depends upon the strength of word syllables emission (Câmara Jr, 1997). Coincidentally, all nouns, verbs and adjectives, and the derivational expansions of these syntactic classes have a stressed syllable (Câmara Jr, 1986), a topic which is not addressed here.

There are still some issues on phonological word that bring some concerns for our study, i.e., its size, since the boundaries of the phonological word are intricate. Actually, there is no consensus among current theorists about the size of a phonological word (Bisol, 2004). The phonological words "mulher" (woman) and "amada" (beloved) → /mu.ˈλɛR/ and /a.ˈma.da/, when pronounced together turn into [mu.ˈλɛ.ra.ˈma.da], by virtue of closed external juncture (sandhi) (Cagliari, 2002), so the resulting phonological words would be /mu.ˈλɛ/ and /ra.ˈma.da/ (Câmara Jr., 1975, Bisol, 2004). Also, sequences like "casa", "de", "pedra" → /ˈka.za/, /di/, /ˈpɛ.dra/, where there is a non-phonological word – the preposition "de" (of) – turn into /ˈka.za di.ˈpɛ.dra/ (stone house). The first example also shows the phenomenon of resyllabification. According to Phonology theory, cases like these are marked by archiphonemes, a category that must be taken into account in the algorithm.

The archiphoneme, a construct elaborated by the Linguistic Circle of Prague is a unit more abstract than the phoneme, resulting from the neutralization of the function to distinguish meanings of one or more traits between two or more phonemes. For example, the archiphonemes |R| and |S| are used in the end of syllables for signaling the neutralization among /s/-/z/, /ʃ/-/ʒ/ and between /r/-/ʁ/, respectively. The pronunciation of these phonemes at the end of words depends on the following word and on the sociolinguistic variety; hence, in the word "folhas" (leaves), the final phoneme is read /s/ if the next sound is unvoiced and the person belongs to the Southern sociolinguistic variety: "folhaspequenas" → [ˈfo.λas pi.ˈke.nas] (small leaves), but it is read /z/ if the next sound is voiced: "folhasgrandes" → [ˈfo.λaz ˈgrɐ̃.dʒis] (big leaves). If the person belongs to the Carioca (Rio de Janeiro) sociolinguistic variety, the same utterances will be read as [ˈfo.λajʃ pi.ˈke.najʃ] and [ˈfo.λajʒ ˈgrɐ̃.dʒiʒ]. Then the archiphoneme is the better solution to register the phenomenon. GonçalvesViana had the intuition to register the archiphoneme |R| by a single grapheme in Portuguese: "r".

Another remarkable concern, which is related to those just discussed, refers to words that are not stressed (clitics), which are usually monosyllabic. Portuguese clitics are only a few (types), but are the most frequently used (tokens), for example, articles, the majority of prepositions, some pronouns and conjunctions and accusative personal pronouns. As has been said, phonologically, clitics are dependent upon the following or preceding stressed word forming with them a prosodic unit. Most Portuguese clitics are monosyllabic, as already mentioned, but there are monosyllabic words that are not clitics, and so are phonological words. In this case, the issue is whether to signal stress on the last ones in the computational program or not. It is important for linguistic research, for example, to make distinctions between stressed and non-stressed monosyllabic words. This topic will be the subject of future research, taking into account the privileges of occurrence of these syntactic classes.

The Portuguese phonological word is a prosodic entity, characterized by a more stressed syllable, preceded or not and followed or not by syllables with two weaker degrees of intensity (Câmara Jr, 1975). The Brazilian linguist Câmara Jr (1986, 1997) has proposed the Theory of Stress Pattern of Brazilian Portuguese (Teoria da PautaAcentual) to explain the subject. The author designed a stress scheme to delimit the phonological word, formalizing all combinatorial possibilities between final and initial syllables of words in relation to their respective distribution of stress. Câmara Jr uses the numbers 3 (primary stress) and 2 (secondary stress) to indicate stressed syllables, 1 for unstressed syllable that comes before the stressed one, and 0 for unstressed syllable that comes after the stressed one.

His theory still needs testing. Other contemporary theories about Brazilian Portuguese stress have been conceived (Bisol, 2004), and we aim to tackle them in the future, for improving the efficiency of the prosodic algorithm, regarding secondary stress, for example. However, for now, our work only deals with the primary stress (3), which will be explained.

The Brazilian Portuguese word prosodic system has been addressed elsewhere, when we dealt with Portuguese spelling system (Vasilévski, 2008, 2012a, 2012b, Vasilévski, Scliar-Cabral & Araújo, 2013), but it is important to return to this issue here, since this is the first time we deal with prosody at the phonological word level specifically. Portuguese is a syllable-stressed language, i.e., the vast majority of Portuguese words has a stressed syllable, leaving aside clitics. However, the stressed syllable is not graphically signaled for the most frequent stressed words (the ones which receive stress on the penultimate syllable and are written with ending "a", "o', "e", followed or not by "s") since Occam's razor principle was adopted, graphically registering only the stress of less frequent stressed words, as already mentioned.

Graphically signaling stress is a powerful hallmark for the reader, because it guides him/her to match the written word with its oral representation in the mental lexicon. The criteria for graphically signaling Portuguese words are the following: a) in which syllable stress falls; b) which is the last vowel, followed or not by "s"; c) which is the last consonant; d) signaling the difference between diphthong and hiatus. The stress diacritics of Portuguese are acute ("lápis" – pencil) and caret ("você" – you). A morpho syntactic diacritic (it does not signal stress) is used for signaling the overlap of the preposition "a" with the definite article or demonstrative pronoun "a(s)" → à(s), or with the same vowel beginning the demonstrative pronoun "aquela(s)", "aquele(s)" → "àquela(s)", "àquele(s)" (Vasilévski, Scliar-Cabral &Araújo, 2013).

In Portuguese, stress may relate to the last, penultimate, antepenultimate or, much more rarely, to the fourth last syllable of the phonological word. The last one may not be considered for European Portuguese, since in Brazilian Portuguese the word "autóctone" (autochthonous) is pronounced [aw.ˈtɔ.kʲ.to.ni] (Vasilévski, 2008), but it is [awtˈɔktunə] in European Portuguese (gɾɐfˈɔnə, 2013) .The stress position clearly reveals the distinctive vowel of a word, but the position of stress does not depend on the phonemic structure of the word (Câmara Jr, 1986). There are no word endings in Portuguese imposing certain stress, but there is a termination which is more frequent, although such frequency is undeterminable phonologically (Câmara Jr, 1997). However, the Portuguese characteristic stress occurs in the penultimate syllable, which gives Portuguese a bass rhythm (Câmara Jr, 1986, 1997). Another characteristic that makes the Portuguese system of signaling the stressed syllable in the written system effective comes from the fact that it was guided by phonological intuition. Portuguese words main stress is graphically registered according to the pattern frequency in the language.

The most frequent word patterns are: …'C(C)V.CV(s)#, …'CV.C(C)V(s)#, …'CV(C).CV(s)# (Scliar-Cabral, 2003, Vasilévski, Scliar-Cabral &Araújo, 2013), where the last vowel must be "a", "e" or "o". This means that the most frequent words in Portuguese are the ones whose stress falls in the penultimate syllable; whose syllables end with vowels; and whose syllables can be open simple (CV) or open complex (CCV), or complex (CVC), and the last syllable can be followed or not by the phoneme /s/. The examples of the most frequent word patterns are:

'CV.CV#: "pato" (duck) → /ˈpa.tu/
'CV.CVs#: "patos" (ducks) → /ˈpa.tuS/
'CCV.CV#: "prato" (dish) → /ˈpra.tu/
'CCV.CVs#: "pratos" (dishes) → /ˈpra.tuS/
'CV.CCV#: "metro" (meter) → /ˈmɛ.tru/
'CV.CCVs#: "metros" (meters) → /ˈmɛ.truS/
'CVC.CV#: "teste" (test) → /ˈtɛS.ti/
'CVC.CVs#: "testes" (tests) → /ˈtɛS.tiS/

These words do not receive any written signal representing stress.

The patterns …C(C)V.'CV(s)#, …CV.'C(C)V(s)# are the second most frequent: the last written vowel must be "á", "é", "ê", "ó", "ô", followed or not by "s" (Scliar-Cabral, 2003, Vasilévski, Scliar-Cabral &Araújo, 2013). In this case, the word stress falls on the last syllable. If the last vowel is [-high, -low], it receives a caret, e.g., "avô" (grandfather) → /a.ˈvo/; if the last vowel is [+low], it receives an acute signal, e.g., "sofá" (sofa) → /so.ˈfa/, "cafés" (coffees) → /ka.ˈfɛS/, "vovó" (grandma) → /vo.ˈvɔ/. On the other hand, if the last stressed vowel is "i" or "u" – for instance, "abacaxi" (pineapple) and "caju" (cashew) → /a.ba.ka.ˈʃi/ and /ka.ˈʒu/ –, the word will not receive any diacritic (Vasilévski, Scliar-Cabral &Araújo, 2013).

Words ending in descending diphthongs without any stress diacritic must be read with stress in the last syllable: "plebeu" (commoner) → /ple.ˈbew/, "união" (union) → /u.ni.ˈãw/. If stress falls in the penultimate syllable in words ending in descending diphthongs, the stressed vowel will be marked with the diacritic: "pônei" (pony) → /ˈpo.nej /. In Portuguese, all words stressed in the antepenultimate syllable, since this pattern is not frequent, have that syllable graphically signaled: "cálida" (warm – fem.) → /ˈka.li.da/, "zênite" (zenith) →/ˈze.ni.ti/, "número" → /ˈnu.me.ru/ (Scliar-Cabral, 2003, Vasilévski, Scliar-Cabral &Araújo, 2013). One example of morphosyntactic function of a diacritic occurs with two verbs – "ter" (to have), "vir" (to come), and their derivatives – in the third person plural, present tense, indicative ("têm", "vêm", "contêm", "provêm"), thus indicating plural, since third person singular is "tem", "vem", "contém", "provém". The pronunciation, however, does not change, since singular and plural forms are homophones: "vem", "vêm"→ /vẽj/, /ˈvẽj/. Therefore, one must understand that the caret in the monosyllabic 3rd person plural is not a prosodic signal but a morphosyntactic one (Scliar-Cabral, 2003).Summarizing, the ending of the word, in its written form, is the criterion that defines the graphic signaling of stress. In written Portuguese, words end with letters that represent oral and nasal vowels, oral and nasal diphthongs. The only consonantal possibilities are the letters "m", "n", "r", "x", "l", "z", "ps":

1) the most frequent words of the language (except clitics), whose stress falls on the penultimate syllable end with the written letters "e", "o", "a", phonemically, /i/, /u/, /a/, followed or not by "s" and do not receive any stress graphic mark;
2) when letters "i" and "u" (followed or not by "s"), phonologically /i/ and /u/, appear at the end of words and the word does not show any stress graphic mark, it means that the word must be pronounced with stress on the last syllable;
3) infrequent patterns of words whose stress falls on the penultimate syllable receive graphic mark when, followed or not by "s", letters "i" and "u", phonologically /i/ and /u/, and descending diphthongs appear at the end of words;
4) All words whose stress falls on the antepenultimate syllable or in the fourth last syllable are coded as receiving graphic stress mark.

These rules and some others covering rare endings work in perfect harmony to graphically reproduce the word prosodic system of Brazilian Portuguese. Regarding the syllable, it is the superior unit in which phonemes (vowels and consonants) combine to allow enunciation (Câmara Jr, 1997). Syllable division is deeply studied by Phonology, and syllable kinds of structure characterize languages. According to Jakobson (Câmara Jr, 1986), the basic phonemic structure is not the phoneme, but the syllable. Hence, it is necessary to address the syllable structure of a language first, in order to make a successful algorithm in recognizing the stressed syllable of a word instead of its stressed vowel. In this way, it is possible to deal with syllable boundaries inside the word and at the end of words, that is, inter-words (external junctures).The syllable in Portuguese can be understood as a set of positions (slope or onset, core or nucleus, and decline or coda) to be occupied by specific phonemes. The nucleus of the syllable is the only essential position in Brazilian Portuguese, and should always be occupied by a vowel, which is the predominant phoneme of the syllable (Câmara Jr, 1986). The nucleus of a non-stressed syllable can be profoundly reduced in Brazilian Portuguese, but is always pronounced. The Brazilian Portuguese syllable was addressed in previous works (Vasilévski, 2012a, 2010, 2011, Vasilévski, Scliar-Cabral &Araújo, 2013) that can be consulted for more details.

The influence of the new orthographic agreement – that will be probably effective in 2016 – in the Brazilian Portuguese word prosody system has been approached before (Vasilévski, 2008, 2012a) as well. The changes do not interfere greatly with the word prosody system, but it is worth remembering that Brazilian Portuguese phonological system as a whole lost some of its transparency, for example, due to the elimination of the diacritic to mark dieresis (trema) and of some differential diacritics. Actually, Brazilian linguists (Perini, 2011) and grammarians (Cunha &Cintra, 2013) do not approve of them. They defend that some diacritics should be preserved for guiding the difference of pronunciation (Cunha &Cintra, 2013), for example, between the preposition "para" (to), a clitic, and the verb "parar" (to stop) in the 3rd person singular of the indicative mode present, i.e, "pára". Another instance, the pronunciation of the verb "aguar" (to water), at the 2nd person singular of the imperative and subjunctive, may be /a.ˈguj/ → "ague" or /ˈa. gwi/ → "águe": in both cases, now there is an ambiguity in reading the final three letters "gue", since they may be read either as /guj/ or as /gi/. The new orthographic agreement makes its pronunciation ambiguous, even for Brazilian Portuguese native speakers, and so does the removal of dieresis.

3. The Word Prosodic Algorithm

3.1. Origin and Architecture

Actually, the algorithm that we call word prosodic is the latest resource developed in the grapheme-phoneme converter Nhenhém (2008-2011) (Vasilévski, 2008, 2012b). The main algorithm of the converter translates texts written in the official spelling of Brazilian Portuguese into phonological characters, and makes them available for using in other applications. As other studies have done (Candeias, Perdigão & Veiga, 2011), we adopted the term phoneme, according to the principles of alphabetic systems, since their aim is precisely to find the best correspondence between phonemes and graphemes (the last ones consisting of one or two letters in Brazilian Portuguese).

Accordingly, other resources of Nhenhémare: phonological statistical report (Vasilévski, 2012a, 2012b), and two syllable parsers (a phonological one and an orthographic one) (Vasilévski, 2012, 2010, 2011), with phonological-syllabic reports (Vasilévski, Scliar-Cabral &Araújo, 2013). In addition, Nhenhém main algorithm was applied to another program, to help speech therapists (Vasilévski 2012b, Blasi & Vasilévski, 2011), and is being applied to a morphological analyzer for Portuguese verbs (Vasilévski, Scliar-Cabral &Araújo, 2012). In addition, it has been used in teaching, in order to help mother language students to distinguish spoken and written language facts, and also in researches of this kind, by students of Linguistics. Moreover, beyond of our projects, Nhenhém grapheme-phoneme conversion rules (Vasilévski, 2008) have proved to be useful for the development of a text-to-speech system with word prosody modeling (Latsch, 2011).

All algorithms have been constructed based on practical linguistic rules. For example, stress marking the vowel of any single word and identifying short contexts in which the correspondence between grapheme and phoneme has stability (Candeias, Perdigão & Veiga, 2011); then marking the syllable boundary and moving the stress mark to the beginning of the correspondent syllable. Regarding the rules for stress assignment, all stress signaling rules (grapheme to phoneme correspondence) seen in section II were inserted into Nhenhém, main algorithm. A brief introduction to the word prosodic algorithm has been presented before (Vasilévski & Araújo, 2013).We were successful in covering the Brazilian Portuguese phonological syllable structure, in part because, from the very first versions, we worked with semivowels, which are essential for establishing syllable boundaries. First, we created a phonological syllable schema (Vasilévski, 2012, Vasilévski, Scliar-Cabral &Araújo, 2013) based on Brazilian linguistic theory (Scliar-Cabral, 2003, Câmara Jr, 1975, 1986, 1997). Such schema, in spite of having some faults, guided us to insert syllable rules into the program. Nhenhém resources undergo constant improvement in order to increase their efficiency. Recently, we adjusted the transcription of the word “muito”, by creating new rules. This word and derivatives perform a pronunciation exception, since the vowels of the syllable “mui” are nasalized. Our first solution was to treat it as an exception and use the symbol /j̃/ for representing the nasalized semivowel. We concluded later that the better transcription is /'mũj.tu/, since, in nasal diphthongs, the nasal feature of the vowel is transferred to the semivowel naturally, with no need to signal its nasality. This simple decision contributed to eliminating an extra phoneme we needed before – j̃/ – and made the system more economic, and coherent, regarding phonemic rules. As we do not test any model or approach, all the restrictions are dealt with by the rules.

Also, we use the archiphonemes |R| and |S|, as mentioned, and this decision makes the program more flexible and able to deal with some unpredictable cases, already detailed (Vasilévski, 2008, Vasilévski, Scliar-Cabral &Araújo, 2013, Vasilévski, 2012b). This choice does not interfere in the word prosodic system of the language; rather, it makes the program more consistent. In any event, the program is capable of replacing the archiphoneme in the cases where the correspondent phonemes are predictable. So, taking into account these two available options, we addressed phonological phenomena that happen in some contexts where a word ends with /s/ or /z/, /r/ or /ʀ/ and the next word starts with a vowel, that is, the mentioned resyllabification phenomenon. As in other programs (Candeias, Perdigão & Veiga, 2011), a problem arises with words derived by affixation, as well as those formed by steem composition, the latter indicated in writing by the use of the hyphen. Nevertheless, we have discussed these points before and solved most of the situations, sometimes by creating morphophonemic rules (Vasilévski, 2012). As said, we have not worked with secondary stress yet. Also, there are some situations raised by some clitics, that have to be edited just now – because we do not use any kind of pre-processing module –, and the program does not put word prosody diacritics in stressed monosyllables that do not have conventional graphic marks. A list or dictionary would work for most clitics, but we believe it is not really necessary to use a dictionary or list for dealing with them.

Thus, we shall start working on rules for covering them (probably morphophonemic rules), as far as it is possible. Before that, it is necessary to evaluate the real influence of clitics and monosyllables in speech transcription.
We also have to work on how the program will deal with this information, so as not to compromise the word prosodic report. As said, it is a complex task. The program uses the symbol ' for signaling stress of vowels and syllables. Marking the stressed vowel did not require the identification of the syllabic unit (Candeias, Perdigão & Veiga, 2011), so this task was performed by Nhenhém since its first version (Vasilévski, 2008); however, the syllable is necessary for researching stress in depth, since the syllable is the basic phonemic structure. Thus, by associating stressed vowel and syllable, we could focus on the stressed syllable, and also the other syllables of the word, and this made it possible to treat some cases of sandhi, which permitted the approach to texts, that is, enunciation, not only words, always using the rules.

More specifically, the proper reproduction of Brazilian Portuguese word prosodic system in the electronic media was only possible after completing some prerequisites, i.e., previous stages of phonemic systematization, which depended upon each other. Firstly, the graphemes were translated into phonemic symbols. Among the phonemes we put the semivowels /j/ and /w/. Then, we worked on the rules to reproduce the phonological phenomena regarding phonemes behavior when combined to form words, that is, short contexts in which the correspondence between graphemes and phonemes is predictable. After that, but also at the same time, since it is not possible to separate these two stages totally, the stress signaling rules were inserted, but it was not possible to work with word prosody properly, because the program marked the stressed vowel, not the stressed syllable. Then, we inserted the syllabic structure rules (orthographic and phonemic) (Vasilévski, 2012, Vasilévski, 2010, Vasilévski, 2011, Vasilévski, Scliar-Cabral &Araújo, 2013), and so it was feasible to approach the stress. The execution steps of the word prosodic algorithm are shown in Figure 1, starting with 'Marcaçãoprosódica', i.e., prosodic marking.

After dividing the input text into phonological syllables, and its stressed syllables being marked, the classification and enumeration of stress marking starts. The program's response to an input text is showed as a syllabic-prosodic-statistical report. The code snippet in Figure 2 is part of the method that generates this procedure.
The method num Tonica (line 240) returns the position of the stressed syllable of the word in variable "s". This process consists of verifying the classification of the position of the stressed syllable, starting from the rightmost syllable to the leftmost one. Thus, the return for the command of the following example ("hidráulico" – hydraulic) will be T3, once the stress mark is in the third leftmost syllable. Between the lines 243 and 267 (Fig. 2) there is the selection structure that classifies the current syllable stress. The unstressed syllables are classified as T0, and the stressed syllable will be classified according to its position in the word (T4, T3, T2 or T1), where T1 refers to the stress on the last syllable of the word, T2 refers to the stress on the penultimate syllable, T3 refers to the stress on the antepenultimate syllable, and T4 refers to the stress on the fourth last one (T refers to intensity type).
The generation of data that will be part of the report is done by means of a storage structure based on a process of feeding lists. The list method was chosen, given its ease of implementation, as a method of dynamic allocation and for having easily accessible count attributes (Nagel et al., 2008).

The program feeds seven lists. Besides being added to its respective lists as rated (T0, T1, T2, T3 or T4), every syllable is also included in a list of general quantitative syllables for the input text (list temp Sil, line 265). From this list, a qualitative list (list temp Sil Dist line, 272) is created, which will be used as reference source for the stressed syllable classification lists, namely, T0, T1, T2, T3 and T4, aiming to verify the stress position for each syllable on the qualitative list. There is a copy of every syllable of the text input in this list, with no repetitions. The result of this procedure is displayed in the word prosodic-statistical report.

3.2. Prosodic-statistical Report

For testing the algorithm, it was applied to the song Construção (Construction), by the Brazilian singer Chico Buarque (1971), for which we obtained the results shown in Table 1.So, this song is composed of 606 syllables, 402 of them are unstressed, and 204 are stressed. There are no words stressed on the fourth last syllable (T4), and 40 words are stressed on the antepenultimate syllable (T3). There are 71 words whose intensity falls on the last syllable (T1), and, as might be expected, most of the stressed syllables of the words, 93, are the penultimate ones (T2). The most common syllable in the song is /si/ (most of them graphically "se", that means, in this specific case, "if"; it is also a syllable in different words, like in "máximo" (maximum) → /ˈma.si.mu/ and "fosse" (verb to be, 3rd person singular of the past subjunctive) → /ˈfo.si/). The syllable /si/ appears 60 times (9.901%), but it is never stressed. The syllable /'ko/ (graphically "co") is the most frequent stressed syllable of the song, as this syllable is present in 24 versus of it.

It always is the first syllable of the word "como", that means "as".By applying the so called prosodic algorithm to a technical text (Construcãosustentavel, 2013), from the area of Building, which contains 617 syllables, that is, 11 syllables more than the amount existing in the song, we obtained the following results:

A comparison between the data obtained from these two discourses leads to some conclusions. For example, the bass rhythm of Portuguese is attenuated in the song, since it has 40 syllables T3. This excess of stressed words on the antepenultimate syllable contributes to the poetic style of the song, and does not happen in the technical text, in which there are only 14 syllables T3. This causes a relevant difference in terms of standard deviation (σ), once, if considered all the syllables of the two texts, we have: σsong = 6.66 and σtechnical text = 3.45. The graph in Figure 3 displays a comparison between the first 42 most common syllables of both texts. The most common syllable of the song, /si/, appears 60 times, and the most frequent syllable of the technical text, /a/, occurs 30 times. The texts follow different patterns until the 11th most frequent syllable, when the occurrences are paired. From this point on, there is no relevant difference between the results, as the curves show. The five most frequent syllables of Portuguese are /a/, /di/, /si/, /ti/ and /u/, in this order ((Vasilévski, Scliar-Cabral &Araújo, 2013), so, in spite of the particular features of these texts, they keep Portuguese syllabic patterns. Soon the program will be capable of properly matching these results with stressed syllables.

The data can be approached from another perspective, like the graph in Figure 4 that shows the stressed syllables distribution for the two texts. We can see that both texts, although belonging to different discourse genres, do not deviate from the stress pattern of Portuguese, given that most of their words have the stress on the penultimate syllable (T2), followed by the words that have the stress on the last syllable (T1), then come the words with stress on the antepenultimate syllable (T3) – the last is rare in Portuguese –, and finally there are the words whose stress falls on the fourth last syllable (T4, "técnica", that means technical, → /ˈtɛ.kʲ.ni.ka/) – this is the rarest stress pattern in Portuguese, as said. It occurs only in the technical text. As mentioned, the last classification is not considered in writing, nor does it apply to European Portuguese.There is much more to be extracted from the data presented, and many interesting issues regarding them arise; however, the main goal here is not to analyze this in detail, but validate the word prosodic algorithm and show how it works.

4. Conclusion

From what has been presented here, we believe that the potential of the word prosodic algorithm to help the investigation of spoken language has been proven, even when using written primary data. A more detailed analysis of the data shown in the report seems promising in the study of word prosody – the intensity stress – of Brazilian Portuguese, addressing various discourse genres. We may argue that it is a useful resource for the description of Portuguese language, taking into account the results obtained from its first version, that is, word prosodic patterns of specific texts and the comparison between them and the general patterns of the language. This makes it possible to envision that, with the improvements that we have been making, the word prosodic algorithm efficiency and, consequently, its usefulness in investigating the language will increase. Then it will be possible to estimate its accuracy rate.An algorithm like this has a value for natural language processing. The results shown by an application of this kind can lead to a better understanding of the word prosodic facts of a language, which may have numerous practical applications, not only for the development of explicit knowledge of the language, but also for the construction of other computational resources that address the oral language.

In this paper, we presented the first version of the word prosodic algorithm, but it is being improved, so as to allow Brazilian Portuguese word stress investigation from different perspectives. In this regard, one of the proposals to be tested soon is the Stress Pattern of Brazilian Portuguese Theory mentioned, proposed by the Brazilian linguist Câmara Jr (1986, 1997). Although dating from the middle of the last century, such a proposition has not yet been properly addressed, mainly due to the lack of technical resources.

This research is supported by the Brazilian government entities CAPES, an agency in charge of promoting high standards for postgraduate courses, and National Council of Technological and Scientific Development (CNPq).

5. References

Buarque, C. Construção, 1971. Retrievedfrom http://letras.mus.br/chico-buarque/45124/.

Bisol, L. (2004). Mattoso Câmara e a Palavra Fonológica. D.E.L.T.A., 20(Esp.), 59-70. Retrieved from http://www.scielo.br/pdf/delta/v20nspe/24261.pdf

Blasi, H., Vasilévski, V. (2011). Programa piloto para transcrição fonética automática na clínica fonoaudiológica. In Documentos para el XVI Congresso Internacional de la ALFAL, Universidad de Alcalá. Alcalá de Henares/Madri.

Cagliari, L. C. (2002). Análise fonológica. São Paulo: Mercado das Letras.

Cagliari, L. C. (2009). A história do alfabeto. São Paulo: Paulistana.

Câmara Jr, J. M. (1975). História e Estrutura da língua portuguesa. Rio de Janeiro: Padrão.

Câmara Jr,J. M. (1986). Estrutura da língua portuguesa. (16th ed.). Petrópolis: Ed. Vozes.

Câmara Jr, J. M. (1997). Problemas de Lingüística descritiva. (16th. ed.) Petrópolis: Ed. Vozes.

Construcão sustentável. (2013). Retrieved from http://www.infoescola.com/ecologia/construcao-sustentavel

Cunha, C.,Cintra, L. (2013). Nova Gramática do Português Contemporâneo, (6th). Lexikon Editorial.

Gonçalves Viana, A. R. (1904). Ortografia nacional: simplificação e uniformização sistemática das ortografias portuguesas. Lisboa: Livr. Ed. Viuva Tavares Cardoso.

gɾɐfˈɔnɘ. (2013). Conversor de Grafemas para Fonemas.Retrieved from http://www.co.it.pt/~labfala/g2p/

Latsch, V. L. (2011). Desenvolvimento de um sistema de conversão texto-fala com modelagem prosódica. PhD-thesis. COPPE. UFRJ: Rio de Janeiro, Brasil.

Malmberg, B. A. (1993). Fonética: teoria e aplicações. Caderno de Estudos Lingüísticos, 1(25), 7-24.

Nagel, C., Evjen, B., Glynn, J., Skinner, S., & Watson, K. (2008). Collections, in Professional C# 2008. pp.250-261. Indianapolis: Wiley Publishing Inc. (Chapter 10).

Parsons, O. A., Pryzwansky, W. B., Weinstein, D. J., & Wiens, A. N. (1995). Taxonomy for psychology. In J. N. Reich, H. Sands, & A. N. Wiens (Eds.), Education and training beyond the doctoral degree: Proceedings of the American Psychological Association National Conference on Postdoctoral Education and Training in Psychology (pp. 45–50). Washington, DC: American Psychological Association.

Perini, M. Interview. (2011). Retrieved from http://www.youtube.com/watch?v=GYobXPh6oRA

Said Ali, M. (1964). Gramática secundária e Gramática histórica da língua portuguesa. (3rd ed.). Brasília: UnB.

Scliar-Cabral, L. (2003). Princípios do sistema alfabético do português do Brasil. São Paulo: Contexto.

The International Phonetic Association. (2013). Retrieved from http://www.langsci.ucl.ac.uk/ipa/

Vasilévski, V. (2008). Construção de um programa computacional para suporte à pesquisa em fonologia do português do Brasil (PhD Linguisticsthesis, Universidade Federal de Santa Catarina, Florianópolis, Brasil). Retrieved from https://repositorio.ufsc.br/bitstream/handle/123456789/91849/254656.pdf.

Vasilévski, V. (2010). Divisão silábica automática de texto escrito baseada em princípios fonológicos. In Anais do III Encontro de Pós-graduação em Letras da UFS (ENPOLE). São Cristóvão, Sergipe, Brasil.

Vasilévski, V. (2011). O hífen na separação silábica automática. Revista do Simpósio de Estudos Lingüísticos e Literários – SELL, 1(3), 657-676.

Vasilévski, V. (2012a). Descodificación automática de lalengua escrita de Brasil basadaenreglas fonológicas. Saarbrücken: Editorial Académica Española.

Vasilévski, V. (2012b) Phonologic Patterns of Brazilian Portuguese: a grapheme to phoneme converter based study. Proceedings of the EACL: Workshop on Computational Models of Language Acquisition and Loss. University of Avignon. France.

Vasilévski, V., Scliar-Cabral, L., Araújo, M. J. (2013). Phonologic and Syllabic Patterns of Brazilian Portuguese extracted from a g2p decoder-parser. International Journal of Advanced Computer Science (IJACSci), 3(8). Retrieved from http://www.ijpg.org/index.php/IJACSci/article/view/469/0

Vasilévski, V., Scliar-Cabral, L., & Araújo, M. J. (2012). Automatic Analysis of Portuguese Verb Morphology: Solving Ambiguities Caused by Thematic Vowel Allomorphs. In Caseli, A. et al (Orgs). Proceedings of the 10th International Conference. PROPOR. (pp. 12-23) Coimbra, Portugal.

Vasilévski, V., & Araújo, M. J. (2013). Um algoritmo prosódico para Português do Brasil. In Anais da III Jornada de Descrição do Português (JDP). Fortaleza, CE, Brazil.

Veiga, A., Candeias, S., Perdigão, F. (2011) Generating a pronounciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment. In Proceedings of the 8thSTIL, SociedadeBrasileira de Computação (pp.144-153). Cuiabá, MT, Brazil.

Tables and figures:

Table 1: Word prosodic data from the song "construção"

	Syllable	Occurr.	T0	T1	T2	T3	T4	%
1	si	60	60	0	0	0	0	9.901
2	mu	30	30	0	0	0	0	4.950
3	a	27	27	0	0	0	0	4.455
4	'ko	24	0	0	24	0	0	3.960
5	ˈfo	22	0	0	22	0	0	3.630
[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]
138	'naw	1	0	0	0	1	0	0.165
	Total	606	402	71	93	40	0	
	%	100	66.34	11.72	15.35	6.60	0.00	100

Table 2: Word prosodic data from the technical text

	Syllable	Occurr.	T0	T1	T2	T3	T4	%
1	a	30	30	0	0	0	0	4.862
2	di	20	20	0	0	0	0	3.241
3	kõ	15	15	0	0	0	0	2.431
4	u	15	15	0	0	0	0	2.431
5	ˈsãw	13	0	13	0	0	0	2.107
	[...]	[...]	[...]	[...]	[...]	[...]	[...]	[...]
228	pu	1	1	0	0	0	0	0.162
	Total	617	444	56	101	14	2	
	%	100	71.96	9.08	16.37	2.27	0.324	100

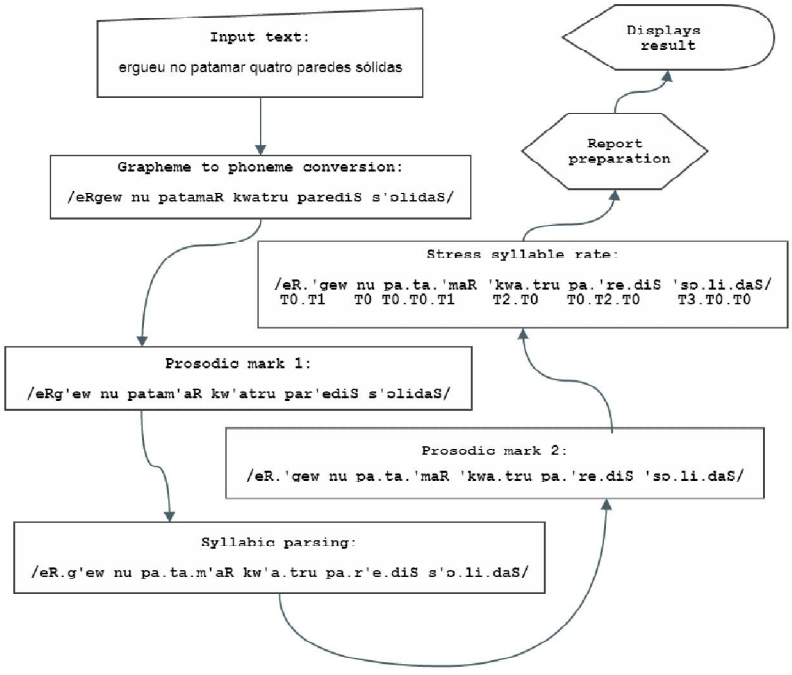

Fig. 1 Flowchart representing the steps of the word prosodic algorithm execution

```
237 ...
238 foreach (string s inpalavras){
239     //Atribuição da classificação da tonicidade silábica
240     tonica = cs.nu mTonica(s);//Retorna T0, T1, T2, T3 ou T4
241     string[] silabas = s.Split('.');
242     i = silabas.Count();
243     foreach (string ss in silabas){
244         #region Adição da silaba em sua classe de tonicidade
245         if ((int)tonica == i) {
246            switch (tonica)
247            {
248                case 1:
249                    T1.Add(ss.Replace("'", ""));
250                    break;
251                case 2:
252                    T2.Add(ss.Replace("'", ""));
253       break;
254                case 3:
255                    T3.Add(ss.Replace("'", ""));
256                    break;
257                case 4:
258                    T4.Add(ss.Replace("'", ""));
259                    break;
260             }
261        }
262        else{
263            T0.Add(ss.Replace("'", ""));
264        }
265        tempSil.Add(ss.Replace("'", ""));
266i        --;
267    }
268 }
269 //Quantidade de sílabas presentes no texto de entrada
270 Total = tempSil.Count;
271 //Filtro das silabas por forma distinta
272 tempSilDist.AddRange(tempSil.Distinct());
273 ...
```

Fig. 2 Code snippet showing how the stressed syllable is selected(Vasilévski&Araújo, 2013)

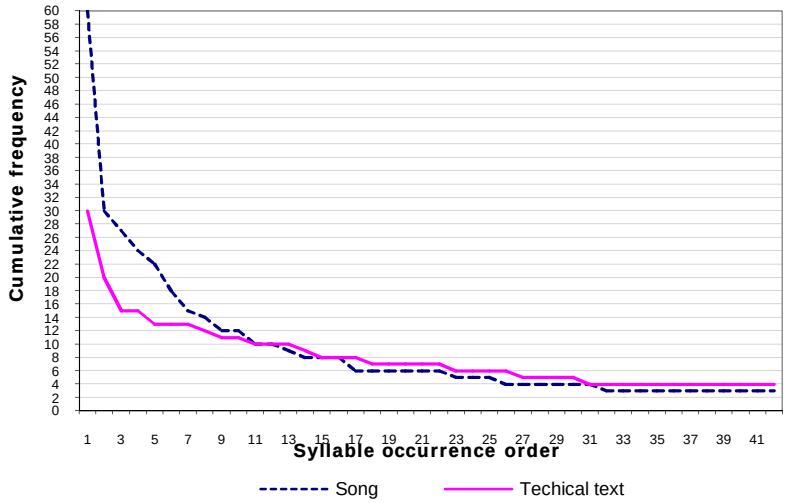

Fig.3Number of occurrence of the 41 most frequent syllables in both texts

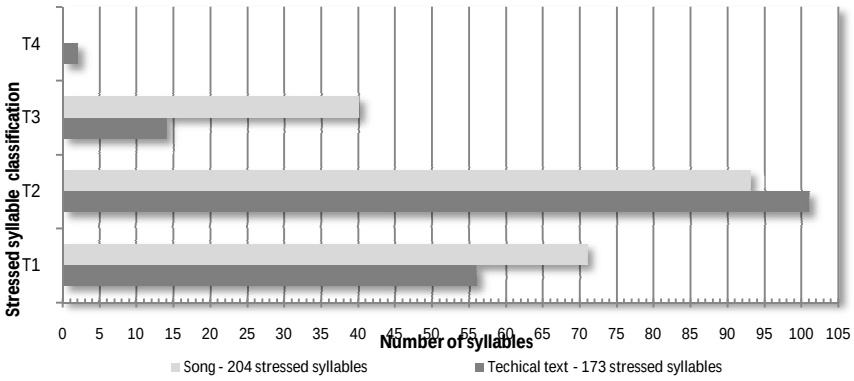

Fig. 4 Syllable distribution of the two texts