AUTOMATIC WORD STRESS MARKER FOR PORTUGUESE TTS

Daniela Braga[1] and Luis Coelho[2]

[1]MLDC – Microsoft Language Development Center, [2]Instituto Politécnico do Porto - ESEIG

ABSTRACT

In this paper, a linguistically rule-based word stress marker for European and Brazilian Portuguese is described. The main goals that led us to develop this application were to increase the grapheme-to-phone performance and to automatically provide lexical stress information to train a Hidden Markov Models based Speech Synthesis System for European Portuguese. The system was implemented and tested giving rise to 99.59% of word accuracy rate for European Portuguese and 99.60% of word accuracy rate for Brazilian Portuguese. This system was also tested with Galician texts and 98.52% of word accuracy rate was obtained.

1. INTRODUCTION

Stress marking has a major impact in two modules of a Text-to-Speech (hereafter TTS) system: on the one hand, in grapheme-to-phone(me) conversion (and syllabification module), and on the other hand in the prosody module. Stress is also part of the phonetic information of the lexicon used as input for the text analysis of a dictionary-based Text-to-Speech system. Stress information used in Hidden Markov Models-based Speech Synthesizers (HTS) offline training has also proved to improve synthetic voice intelligibility and naturalness [1]. Although word stress in Portuguese is widely studied in literature [2], there is not much work published on automatic word stress marking for Portuguese. In the early 90's, Oliveira et al. [3] pointed out the importance of stress marking and refer that the DIXI version uses 18 rules. More recently, Teixeira et al. [4] described a stress marker algorithm with only 3 rules, followed by a table of exceptions (although not published), while Barros & Weiss [5] presented a maximum entropy-based stress model which was trained with a 4219 word stress annotated corpus. In [4], no performance rates of this rule-based system are presented. In [5], the accuracy rate of the proposed statistical method is 85.57%. In this paper, we present a tool for automatic word stress marking for European and Brazilian Portuguese. This tool represents the latest development of a preliminary version presented in a previous work [6], which was designed only for Brazilian Portuguese language and whose results were 98.58% of accuracy rate. Our current version has not only largely overcome the initial performance results, but it has also turned to be more flexible, supporting both European and Brazilian Portuguese varieties. This paper is structured as following: in section 2, the rule-based automatic word stress marker for Portuguese is presented; in section 3, the tests are described and the results are discussed; in section 4, the application of this work to Galician is presented and discussed; in section 5, main conclusions are summarized and future work is foreseen.

2. AUTOMATIC WORD STRESS MARKER

The proposed word stress marker is composed by 31 rules and is based on the analysis of context around the last graphemes of each word. After the text is separated into sentences and the sentences are separated into words, the system checks word by word in order to find non stressed words, which are pre-defined and receive no stress mark. According to literature [7], non stressed words are monosyllabic high frequent function words such as monosyllabic definite and indefinite articles (<o, a, os, as, um, uns>); clitics (<me, te, se, o, a, os, as, lo, la, los, las, no, na, nos, nas, lhe, lhes, nos, vos>) and their contractions (<mo, ma, mos, mas, to, ta, tos, tas, lho, lha, lhos, lhas, no-lo, no-la, no-los, no-las, vo-lo, vo-la, vo-los, vo-las>); relative pronoun <que>; monosyllabic prepositions (<a, com, de, em, por, sem, sob>) and their contractions (<do, da, dos, das, ao, à, aos, às, no, na, nos, nas, num, nuns>); and monosyllabic conjunctions (<e, mas, nem, ou, que, se>). The last grapheme of the word, which receives the position number zero ^(0), is the starting point for each rule. Then, the left context of this zero position is analyzed grapheme by grapheme and the stressed vowel is predicted according to the different combinations of the graphic patterns. The symbol set used in the rules design is shown in Table 1. In Table 2, the word stress marker complete algorithm is displayed. Most of the rules are repeated (e.g. rules 5 and 6, rules 7 and 8, etc.), bearing in mind the adjustment of the graphemes' position, so as to predict plurals of nouns and adjectives. The stressed vowel is marked with a digit (<1>) and not with an apostrophe, in order to avoid ambiguities in the TTS text normalization module.

symbol	Meaning
^(0)	Word last grapheme
^(1)	Word penultimate grapheme
^(2)	Word antepenultimate grapheme
^(3)	Word third last grapheme

^(4)	Word fourth last grapheme
T	Position occupied by the stressed vowel
T=1	Stressed vowel is the penultimate grapheme
/	Except
→	Then
{x}	Grapheme x
{ }	Space
1	Stressed vowel

Table 1. *Symbols used in the automatic word stress marker algorithm for EP and BP.*

#	Rule	Exemple
1	List of non stressed words.	por, um, se
2	If there is an orthographic accent[1], the accented vowel is the stressed one. The acute or circumflex accents have precedence over the tilde[2].	órgão, órgãos, bênção, bênçãos
3	If the word has only one vowel → T= vowel	tem, vem, bem, vi
4	If ˆ(0) = {r, l, z, x} → T = 1	propor, juiz
5	If ˆ(0) = {m} and ˆ(1) = {i,o,u} → T = 1	pudim, bombom, comum
6	If ˆ(0) = {s} and ˆ(1) = {n} and ˆ(2) = {i,o,u} → T = 2	pudins, comuns
7	If ˆ(0) = {i} and ˆ(1) = {u,} and ˆ(2) = {q,g} → T = 0	caqui, aqui, sagüi
8	If ˆ(0) = {s} and ˆ(1) = {i} and ˆ(2) = {u,} and ˆ(3) = {q,g} → T = 1	caquis, sagüis
9	If ˆ(0) = {i,u} and ˆ(1) é vogal → T = 1	caju, grau, pneu
10	If ˆ(0) = {i,u} and ˆ(1) is not a vowel → T = 0	caju, javali
11	If ˆ(0) = {s} and ˆ(1) = {i,u} and ˆ(2) is not a vowel → T = 1	cajus, javalis
12	If ˆ(0) = {s} and ˆ(1) = {i,u} and ˆ(2) is a vowel → T = 2	andais, pauis, graus.
13	If ˆ(0) = {and} and ˆ(1) = {u} and ˆ(2) = {q, g} and ^(3) is vowel/{u} → T = 3	Alambique, Henrique, obrigue
14	If ˆ(0) = {s} and ˆ(1) = {e} and ˆ(2) = {u} and ˆ(3) = {q, g } and ^(4) is vowel/{u} → T = 4	alambiques, Henriques, obrigues
15	If ˆ(0) = {e} and ˆ(1) = {u} and ˆ(2) = {q, g} and ^(3)={u} → T = 4	açougue, azougue, tougue
16	If ˆ(0) = {s} and ˆ(1) = {e} and ˆ(2) = {u} and ˆ(3) = {q, g} and ^(4)={u} → T = 5	açougues, azougues, tougues
17	If ˆ(0) = {e} and ˆ(1) = {u} and ˆ(2) = {q, g } and ^(3) ={r} → T = 4	embarque, marque, morgue
18	If ˆ(0) = {s} and ˆ(1) = {e} and ˆ(2) = {u} and ˆ(3) = {q, g } and ^(4) ={r} → T = 5	embarques, marques, morgues
19	If ^(0) = {e} and ^(1) = {u} and ^(2) = {q, g} and ^(3) ={n}→ T = 4	sangue, , manque,
20	If ^(0) = {s} and ^(1) = {e} and ^(2) = {u} and ^(3) = {q, g} and ^(4) = {n} → T = 5	exangues, manques, palanques
21	If ˆ(0), ˆ(1), ˆ(2) are vowels, if ˆ(1) = {i,u} and if ^(3) is a consonant, { } → T = 2	meia, seio, apoio, aia, gaia, papagaio
22	If ^(0) = {s, m} e ˆ(1), ˆ(2), ˆ(3) are vowels, if ˆ(2) = {i,u} and if ^(4) is a consonant, { } → T = 3	meias, seios, gaias, papagaios
23	If ^(0) and ^(3) are vowels, and ^(1) is a consonant and ^(2) = {i, u} and ^(4) ≠ vowel/ {u} → T = 3	cadeira, queima, louco, estrangeiro
24	If ^(0) = {s} and ^(1) and ^(4) are vowels, and ^(2) is consonant and ^(3) = {i, u} and ^(5) ≠ vowel/{u} → T = 4	cadeiras, queimas, loucos, estrangeiros
25	If ˆ(0) = {a,e,o} and ˆ(1) is consonant and ˆ(2) = {n} and ˆ(3) = {i,u} and ˆ(4) is vowel→ T = 3	ainda, caindo, incluindo, oriundo
26	If ˆ(0) = {s} and ^(1) = {a,e,o} and ˆ(2) is consonant and ˆ(3) = {n} and ˆ(4) = {i,u} and ˆ(5) is vowel→ T = 4	Oriundos
27	If ˆ(k)[3] = penultimate vowel and ˆ(k) = {i,u} and ˆ(k+1) is a vowel and ˆ(k-1) is not a vowel and ˆ(k+2) is not {q,g} → T = k+1	outro, claustro
28	If ˆ(0) = {m} and ˆ(1) = {e} and ˆ(2) = {u} and ˆ(3) = {q} → T = 1	Quem
29	If ^(0) = {a,o, e} and ^(1) = {i, u} and ^(2) is a cons. or {u} → T = 1	inicie, assobio, continua, rua
30	If ^(0) = {s, m} and ^(1) = {a,o, e} and ^(2) = {i, u} and ^(3) is a consonant or{u} → T = 2	academias, continuam, inicием
31	If none of the above rules occur → T = penultimate vowel of the word	casa, homem, guerra

Table 2. *Rule set for word stress marker in European and Brazilian Portuguese.*

3. IMPLEMENTATION, TESTS AND RESULTS

The word stress marker was programmed in C/C++ for Windows. A graphic interface was built in Borland Delphi with the purpose of testing the performance of this application. The stress output is combined with the syllabification output, as can be seen in Figure 1. The automatic syllabification application shown in Figure 1 was already described in [8]. Syllabification and stress prediction make stress information be envisaged in a syllable unit context and not only in a vowel context. This way, vocalic stress prediction can be extended to syllabic stress prediction. The word stress marker is part of a larger application which is basically the Portuguese HTS front-end, presented in [9]. Two tests were

[1] Exception to this rule: in the following words <àquele, àqueles, àquela, àquelas, àqueloutro, àqueloutra, àqueloutros, àqueloutras>, the orthographic accent <`> should not be considered as an accent. This is an accent that marks the contraction between two words, not a phonological stress.

[2] Exceptions to this rule occur in words ending by the suffixes <-inho>,<- inha>,<- inhos>,<- inhas>,<-zinho>, <-zinha>,<-zinhas>,<- zinhos> (e.g. pãezinhos, sotãozinho) or <-mente> (e.g. cristãmente), in which the stressed vowel becomes the penultimate syllable, although the accented vowel still has a secondary accent.

[3] (k) is a variable, a given grapheme.

conducted in order to assess the performance of the automatic word stress marker with corpora from both varieties of Portuguese (European and Brazilian). The first test was carried out using 1000 sentences as input, randomly extracted from Cetem-Público European Portuguese (EP) newspaper corpus [10], containing 8052 words and 41156 characters without spaces. The second test was conducted using 500 sentences, randomly extracted from Ceten-Folha Brazilian Portuguese newspaper corpus [11] and composed by 5372 words and 28633 characters without spaces. The overall results show a very similar performance of the word stress marker, giving rise to 99.59% of accuracy rate for European Portuguese and 99.60% of accuracy rates to Brazilian Portuguese. Table 3 shows the results of the word stress marker using Cetem-Público European Portuguese corpus. The word error rate (hereafter WER) is 0.41%, from which 0.35% occur in foreign words. In fact, foreign words are the main cause for errors in our system. In Figure 2, a detailed display of errors according to their origin can be seen. English origin words in Portuguese language have the highest percentage of errors (0.11%), as shown in Figure 2, and occur in words such as <internet> (in_te1r_net) or <cocktail> (coc_kta_i1l). This result is explained by the high frequency of English origin words in the Portuguese vocabulary.

Type of error	# errors	% errors
Foreign words	28	0.35
Portuguese words	5	0.06
Total	**33**	**0.41**

Table 3. *Results of word stress marker using European Portuguese corpora.*

With the same number of errors (0.11%), we can find foreign words from other origins, of which <jihad> (ji1_had) or <Arafat> (a_ra1_fat) are examples. Italian, French and Latin origin words are responsible for 0.04% of errors each, and can be found in brands <Lamborghini> (lam_bor_ghi_ni1), proper names <Pausini> (pau_si_ni1), <Jacques> (jac_qu1_es), <Chirac> (chi1_rac) and compound expressions <ex-libris> (e1x li_bri1s). The Portuguese words' category includes errors in stress prediction of the words <Coimbra> (co1im_bra) and <Quercus> (quer_cu1s), which were repeated several times in the corpus. The comparison of our results with other accuracy rates reported in literature using statistical methods (85.57% stated in [5]) seems to demonstrate the better performance of a linguistically rule-based approach when tackling the word stress prediction. In Table 4, the results of the word stress marker using Ceten-Folha Brazilian Portuguese corpora are presented. Once more, foreign words such as <Corinthians> (co_rin_thi1_ans) are the major cause for the system WER (being responsible for 0.31% of the errors). The word <porque> is the second error cause. Although in European Portuguese this word is stressed in the second last syllable (<porque>), in Brazilian Portuguese this word is stressed in the last syllable (<porque>). This difference will be included in the system and treated as an exception. One error occurred in a readable acronym <Telesp> (te1_lesp), because it shows an unpredicted final graphic pattern. These results represent a great improvement when compared with others, previously described in literature (98.58% of accuracy rate in [6]).

Type of error	# errors	% errors
Foreign words	17	0.31
<porque>	4	0.07
Acronyms	1	0.01
Total	**22**	**0.40**

Table 4. *Results of word stress marker using Brazilian Portuguese corpora.*

4. APPLICATIONS TO GALICIAN

The common historical origin between Portuguese and Galician and their linguistic proximity led us to test the proposed word stress marker with Galician corpora. The selected corpus was composed by 300 sentences, 2627 words and 12250 characters without spaces, randomly extracted from CORGA - Corpus de Referencia do Galego Actual [12]. This corpus is a collection of different sources (oral and written) and genders (newspapers, literature, magazines, etc.). No adaptation of the here described word stress marker was made to Galician language, except in the non stressed words' list. The analysis of the results using Galician texts was based in the requirements presented in [13]. In Table 5, it can be seen that the accuracy rate of the automatic Portuguese word stress marker when tested with Galician texts is 98.52%. This encouraging result not only demonstrates a very similar phonological structure between Portuguese and Galician, but also proves the high applicability of this module to a different romance language without any algorithm adaptation.

Type of error	# errors	% errors
For lack of accent in Galician	34	1.29
Foreign words	2	0.08
Others	3	0.19
Total	**39**	**1.48**

Table 5. *Results of word stress marker for Galician.*

Most of the errors shown in Table 5 are due to the fact that in Galician, because of the Spanish orthography influence, there is no graphical accent in words ending with diphthongs like /jo/ or /ja/ (e.g. <contrario, media, principio, Emilio, circunstancia>), because these words are considered to be stressed in the penultimate syllable. However, the same words exist in Portuguese but are considered to be stressed in the antepenultimate syllable, which means that the Galician diphthongs are considered to be two syllables in fact. Hence, these

words in Portuguese receive a graphical accent (e.g. <contrário, média, princípio, Emílio, circunstância>). Therefore, the high rate of errors (1.29%) in Table 5 can be explained because the graphical accent is essential to the identification of the tonic syllable and Galician doesn't have it in the same contexts as Portuguese does. Anyway, according to [14], there is a trend in Galician oral language to pronounce these words like in Portuguese, in other words, separating these final diphthongs in two syllables. In order to solve these errors in a great extent we could propose the following rules, as shown in Table 6:

#	Rule	Example
1	If ^(0) = {a,o} and ^(1) = {i} and ^(2) is consonant and ^(3) = V → T = 3	contrario, media, principio
2	If ^(0) = {s} and ^(1) = {a,o} and ^(2) = {i} and ^(3) is consonant and ^(4) = V → T = 4	contrarios, medias, principios
3	If ^(0) = {a,o} and ^(1) = {i} and ^(2) is consonant and ^(3) = {m,n} and ^(4) = V → T = 4	circunstancia
4	If ^(0) = {s} and ^(1) = {a,o} and ^(2) = {i} e ^(3) is consonant and ^(4) = {m,n} and ^(5) = V → T = 5	circunstancias

Table 6. *Stress marking rules to solve words ending by diphthongs /jo/ and /ja/ in Galician.*

The proposed rules in Table 6 would be able to raise the current accuracy rate to 99.81%. Other errors occur in foreign words (0.08%), similarly to Portuguese, in words like <Madrid> (ma1_drid) and <chofer> (cho_fe1r) and in abbreviations, such as <mili (abbreviation of "servicio militar")> (mi_li1). These encouraging results and small refinements allow us to conclude that our system and approach are highly applicable to other languages in general and to romance languages in particular.

5. CONCLUSIONS

In this article, a linguistically rule-based automatic word stress marker for European and Brazilian Portuguese was described, implemented and tested. The purpose of this work was to provide stress information to the front-end part of the TTS system (syllable boundary marker and grapheme-to-phone(me) transcriber) and to the training corpora used by the HTS back-end, since it was proved in [1] that this information improves synthetic naturalness. The proposed automatic word stress marker deals with 31 rules and starts analyzing the last grapheme of a word. The goal is to identify the tonic vowel of each word. Combined with the automatic syllabification information, the stress information can affect the entire syllable and not only the stressed vowel. This approach proved to be very efficient giving rise to very encouraging accuracy rates when tested with real text corpora: 99.59% with European Portuguese corpora and 99.60% with Brazilian Portuguese corpora. This system was also experimented with Galician corpora with a small adaptation in the non stressed word list, giving rise to 98.52% of accuracy rate. A refinement of these results based on the errors' analysis was proposed. Due to the success of application of the system presented in this paper to European Portuguese, Brazilian Portuguese and Galician, we believe that this approach can be easily adapted to other languages. The application of this work to Catalan was already done with similar success [15] and other languages are envisaged.

6. REFERENCES

[1] Maia, R.: Speech Synthesis and Phonetic Vocoding for Brazilian Portuguese Based on Parameter Generation form Hidden Markov Models. PhD thesis. Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan (2006)

[2] Mateus, M., Andrade, E. *The Phonology of Portuguese*. Oxford University Press, Oxford, 2000.

[3] Oliveira, L., Viana, M., Trancoso, I."DIXI - Portuguese Text-to-Speech System", Proceedings of EUROSPEECH'91 - 2nd European Conference on Speech Communication and Technology, pp.1239-1242. Genoa, Italy, 1991.

[4] Teixeira, J. P., Freitas, D."MULTIVOX- Conversor Texto-Fala para Português", Lima, V. (eds.) III Encontro para o Processamento Computacional do Português Escrito e Falado (PROPOR'98), pp. 88-98. Porto Alegre, RS, Brazil, 1998.

[5] Barros, M., Weiss, C. "Maximum Entropy Motivated Grapheme-to-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech", IV Jornadas en Tecnologías del Habla, pp. 177-182. Zaragoza, España, 2006.

[6] Silva, D., Lima, A., Maia, R., Braga, D., Moraes, J. F., Moraes, J. A., Resende Jr., F. "A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing", VI International Telecommunications Symposium (ITS2006), pp.550-554. Fortaleza-CE, Brazil, 2006.

[7] Cunha, C., Cintra, L. *Nova gramática do português contemporâneo*. Sá da Costa, Lisboa, 1992.

[8] Braga, D., Resende Jr., F. G. V.: "Módulos de Processamento de Texto Baseados em Regras para Sistemas de Conversão Texto-Fala em Português Europeu", XXI Encontro da Associação Portuguesa de Linguística, pp.141-156. Coimbra, Portugal, 2007.

[9] Braga, D. *Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português*. PhD Thesis. Universidade da Coruña, España, 2008.

[10] Cetem-Público,http://www.linguateca.pt/CETEMPublico/

[11] Ceten-Folha, http://www.linguateca.pt/CETENFolha/

[12] Corpus de Referencia do Galego Actual, http://corpus.cirp.es/corga/

[13] Real Academia Galega/ Instituto da Língua Galega: Normas ortográficas e morfolóxicas do idioma galego. Real Academia Galega/ Instituto da Língua Galega, Vigo, España, 2003.

[14] Freixeiro Mato, X. R. *Manual de Gramática Galega*. Edicións a Nosa Terra, Vigo, 2006.

[15] Rustullet, S.; Braga, D.; Nogueira, J.; Dias, M. "Automatic Word Stress Marking and Syllabification for Catalan TTS", Proceedings of Interspeech 2008, Brisbane, Australia, 2008.