Automatic Disambiguation of Homographic Heterophone Pairs Containing Open and Closed Mid Vowels

Christopher Shulby[1], Gustavo Mendonça[1], Vanessa Marquiafável[2]

[1]Instituto de Ciências Matemáticas e de Computação –
Universidade de São Paulo (USP)

[2]Instituto de Biociências, Letras e Ciências Exatas –
Universidade Estadual Paulista (UNESP)

{chrisshulby,gustavoauma}@gmail.com, marquiafavel@yahoo.com.br

Abstract. *The issue of openness in Brazilian Portuguese vowels is a question not yet satisfactorily explored in the field of automatic classification of homographic heterophones (HH). Therefore, we aimed to develop and test a pilot classifier which assists in the automatic disambiguation of HH. For this purpose, a set of 226 word pairs of HH with the unique grammatical classes, distinguished by alternating mid vowels [e, E] and [o, O], was analyzed. The results showed that the rules proposed herein solve most disambiguation problems of HH word pairs containing mid vowels in the corpus analyzed and can be applied to TTS and ASR applications. The data also revealed that a predominant trend of non-verb classes exists, and, for some word pairs, that value can reach 95% occurrence.*

1. Introduction

Speech is the most simple and natural form of human communication. Automatic speech processing technology is relatively recent and still enjoying its youth. 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. This study aimed to bring contributions to the field of automatic speech processing of Brazilian Portuguese, whether for the linguistic studies that underpin it or for the software industry focused on NLP (Natural Language Processing). The objective was to conduct a case study, based on Brazilian Portuguese (hereafter BP) regarding the automatic disambiguation of homographic heterophone pairs (hereafter HH): pairs of words with the same spelling but unique pronunciations. In the present study, we analyzed HH pairs with distinct grammatical classes which differ in pronunciation, consisting of mid vowels [e, E] and [o, O], for example: inter[e]sse (noun) ~ inter[E]sse (verb); s[o]bre (preposition) ~ s[O]bre (verb); and j[o]go (noun) ~ j[O]go (verb). The principal difficulty lies in determining the transcription of <e> and <o> typed mid vowels when they contain a stressed syllable within a word that is not marked by an accent, due to the fact that it can be pronounced at times either as an open [E] or [O], while other times as the closed [e] or [o] varieties.

Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, pages 126–137,
Fortaleza, CE, Brazil, October 21–23, 2013.

Portuguese orthography is phonographic and alphabetic in nature. The graphic symbols that are used in writing BP seek to represent sounds that occur in speech [Robinson 2006]. Although alphabetic systems are those in which greater similarity between speech and orthography can be observed, there is no pure natural system. In other words, no orthographic system with natural origins in a speech community has a perfectly bi-univocal relationship between phone(me) and grapheme. This is due to the tensions between the dynamism of phonetic evolution within all living languages and the conservative nature, typical of all orthographies. This type of non-isomorphism between graphemes and phone(me)s in BP causes difficulties in situations where it is necessary to convert grapheme to phone(me) and vice versa. An example of this can be noted in writing acquisition (by both native and non-native speakers): the lack of parity between phone(me)s and graphemes is problematic and often, spelling errors committed by students arise due to transcription generalizations while writing a word based on how it is pronounced [Miranda 2006]. Another example of these difficulties can be seen with the development of automatic speech synthesis and recognition systems. In view of the non-mutual unambiguity between graphemes and phone(me)s, the tasks of synthesizing or recognizing speech can be regarded as quite complex, since the direct mapping between the sound signal and the written representation of a particular word or phrase does not necessarily match, adjustments must be made where the relationship between grapheme and phone(me) is not explicit.

HH disambiguation is of great importance, even if the amount of HHs existing in an excerpt represent a very small number of instances in relation to the whole. In the context of speech synthesis, when an inappropriate sound (not agreeable/acceptable or even unintelligible to the listener/user of a system) is played, it ends up attracting an unnecessarily great deal of user attention to the error. Although the error rate could almost be considered negligible compared to the system's success, the user will tend to question and often evaluate the efficiency of the synthesis system inadequately. Given its importance, HH research has been widely explored by the scientific community for many languages including English, Japanese, Chinese and Thai.

In the case of European Portuguese (EP), the problems of converting graphemes and phone(me)s emerge mainly in the transcription of stressed mid vowels [e, E] and [o, O] [Veiga et al. 2011]. In that study, it was found that over 80% of conversion errors were caused by these exact same mid vowel shifts in stressed syllables. Mid vowels involve a range of linguistic phenomena in BP as they: (i) have different status [Miranda 2006], (ii) are realized in the context of rising pre-stressed syllables [Bortoni et al. 1992], [Viegas 2006], (iii) are metaphonic in nouns and verbs [Tomaz 2006] and (iv) shift vowel spaces [Roces 2010]. In this paper, we discuss a specific case of problems involving mid vowels: the automatic disambiguation of HH word pairs, that belong to different grammatical classes and are distinguished by openness of a mid vowel in stressed syllables (e.g., inter[e]sse (noun) ~ inter[E]sse (verb); and s[o]bre (preposition) ~ s[O]bre (verb)). With respect to stressed mid vowels [e, E] and [o, O] HH pairs, disambiguation is alike whether in BP or EP. Undoubtedly, language usage is different between BP and EP, but the linguistic

structural context that determines mid vowel stressed HH pair disambiguation is the same. Previous studies have discussed this problem for EP [Braga and Marques 2007; Braga 2008], using a rule-based algorithm for a minor group of HH pairs with mid vowels. For BP, Silva et al. (2009) employed a rule-based method which uses morphosyntactic information from word libraries built specifically for disambiguation. Manual classification methods were also put to use for BP [Cristófaro-Silva 2005].

In this paper, we propose a disambiguation method which addresses an extensive number of HH pairs with mid vowels, by using two contextual rules and morphosyntactic information from the automatic tagger, MXPOST. It should be noted that smaller sets of such words exist which must be treated by other methods. It should be reinforced that this paper only explores the large number of HH pairs with mid vowels which can be disambiguated by PoS tags alone.

2. Methodology

2.1 Selection and classification of HH word pairs

Initially, we collected HH pairs present in two BP dictionaries published on the topic, namely: *O Dicionário de Palavras Homógrafas*, by Walmírio de Macedo (1961); and *O Novo Dicionário de Acentuação das Palavras Homógrafas Heterófonas*, by Pandiá Pandu (1972), a total of 1,812 pairs were collected. Of these, 772 repeated pairs (present in both dictionaries) as well as 141 containing the same grammatical category were excluded; therefore, 899 HH pairs were initially considered. Next, we verified the occurrence of the pairs considered in the MAC-Morpho [Aluísio et al. 2003][1] corpus. The choice of MAC-Morpho was due to several reasons: first, it is a corpus of BP, which contains 1,167,183 words from newspaper articles taken from the *Folha de São Paul*o in 1994 and, more specifically, from the following 10 sections: *Agronomia (ag), Brasil (br), Cotidiano (co), Dinheiro (di), Esporte(es), Ciência (fc), Informática (if), Ilustrada (il), Mais! (ma) e Mundo (mu)*. Secondly, the MAC-Morpho has morphosyntactic annotation: initially, the corpus was annotated automatically by the parser *"Palavras"* [Bick 2000]; then, was reviewed manually by linguists. Thirdly, the corpus is freely available to the public on the web. After a thorough automatic analysis, 226 of the 899 HH pairs also appeared in MAC-Morpho. This small number of pairs could be due to the discrepancies which exist between the outdated dictionaries, containing all but dead; archaic words, and the modern lexicon (specifically in a journalistic setting). The 673 remaining pairs, registered in the dictionaries but which did not occur in the MAC-Morpho, were disregarded in the analysis due to the lack of reliable annotations.

The next step consisted of a typological pair analysis, according to its oppositional grammatical nature and phonetic alternation (openness) present in varieties of the stressed

1 The MAC-Morpho corpus is available in the *Lácio-Web* project, on the *Núcleo Interinstitucional de Linguística Computacional* (NILC) website: http://www.nilc.icmc.usp.br/lacioweb/.

mid vowels <e> and <o>. All word pairs were divided into four groups (types) as specified in Table 1.

Table 1. HH pair classification

Homographic Heterophone pair classification		
Class Number	**Type of oposition**	**Frequency**
1	[E] verb/ [e] noun or adjective	100
2	[E] verb / [e] other parts of speech (non-nominal)	12
3	[O] verb/ [o] noun or adjective	102
4	[O] verb / [o] other parts of speech (non-nominal)	12
	Total:	226

2.2 Rules used for HH pair disambiguation

In order to disambiguate the 226 word pairs, two rules, derived from their parts of speech (for our purposes only nouns and verbs), were considered. These two rules selected for demonstration in this study were chosen because they are responsible for the disambiguation of most possible HH pairs found in BP. The first rule states that if the vowel <e> is contained in the pair's stressed syllable and it is classified as a noun, it must be transcribed as [e]; thus, a closed-mid vowel. Otherwise, the <e> should be transcribed as [E], a front, open-mid vowel classified as a verb. Conversely, the second rule states that if the vowel <o> is present in this same circumstance and the word is classified as a noun, then <o> should be transcribed as [o], a back, closed mid-vowel. Otherwise, it should be [O], a back, open-mid vowel, classified as a verb. Both rules apply similarly to other pair groups classified in section 2.1, as there is always a shift in grammatical class by virtue of openness of mid stressed vowels in BP.

It is also worth mentioning that the creation of disambiguation rules will be essential in the authors' graduate projects, which include intelligent tutor systems and a grapheme-phone converter. Our aim with this work is primarily to examine the relationship of openness and grammatical class present in HH pairs containing mid stressed vowels, as well as to evaluate the actual accuracy of these disambiguation rules created for the pairs listed in the two BP dictionaries, with application to a real corpus (in this case MAC-Morpho), based on information coming from the word classes as a result of the automatic morphosyntactic tagging by MXPOST. More details about the MXPOST tagger are given in the next section.

2.3 The morphosyntactic tagger MXPOST

MXPOST[2] is an automatic morphosyntactic tagger, based on the method of maximum entropy, which was originally developed for English by Ratnaparkhi (1996) and adapted for

2 MXPOST can be downloaded from the NILC's website: http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html.

BP by Aires et al. (2000). MXPOST achieves high accuracy in assigning words to classes of lexical items: 97% [NILC 2000] and is freely available on the web. MXPOST was used to automatically tag the MAC-Morpho corpus, then the manual and the automatic word class tags were compared and decisions were made based on agreement, maintaining the MAC-Morpho manual annotations as the gold-standard. From the data obtained by word class, a classifier was developed and piloted to demonstrate how a PoS tagger can be applied to HH pair disambiguation. The aim of the classifier was to automatically categorize HH pairs by their part of speech labels. In order to do this, we investigated the alternatives currently known to analyze spoken input, such as software programs for speech recognition [Silva et al. 2010], morphosyntactic taggers such as MXPOST [Aluisio et al. 2003] and grapheme-phoneme converters [Frunza and Inkpen 2009], which may be adapted to the system in order to achieve the best results.

3. Results

After the data analysis, it was found that most HH pairs with differences in their vowels (open or closed) can be differentiated by their parts of speech alone, based on the application of the two rules explained in detail above. For the sake of space these rules will not be repeated here.

First of all, we removed the morphosyntactic labels present in the MAC-Morpho corpus, to obtain the "raw" texts. Next, we used the morphosyntactic tagger, MXPOST [Ratnaparkhi 1996], to automatically annotate those texts, in order to verify the accuracy with respect to HH pairs between the automatically and manually annotated versions of MAC-Morpho, we prepared a PHP script to extract the 226 HH occurrences from both corpora. In total, 226 types of HHs accounted for 16,851 occurrences. It was found that from those 16,851 instances, the correlation between the automatic and manual annotations of MAC-Morpho obtained 90.1% agreement. In order to properly execute this task, it was necessary to adapt the tagsets, given that the MXPOST and MAC-Morpho annotation schemas differ in the number and type of word classes that they consider. For example, MAC-Morpho includes the classes "v" and "vaux" as verbs, whereas MXPOST only uses the class "verb". The following chart illustrates the success rate of MXPOST regarding HH pairs by frequency groups:

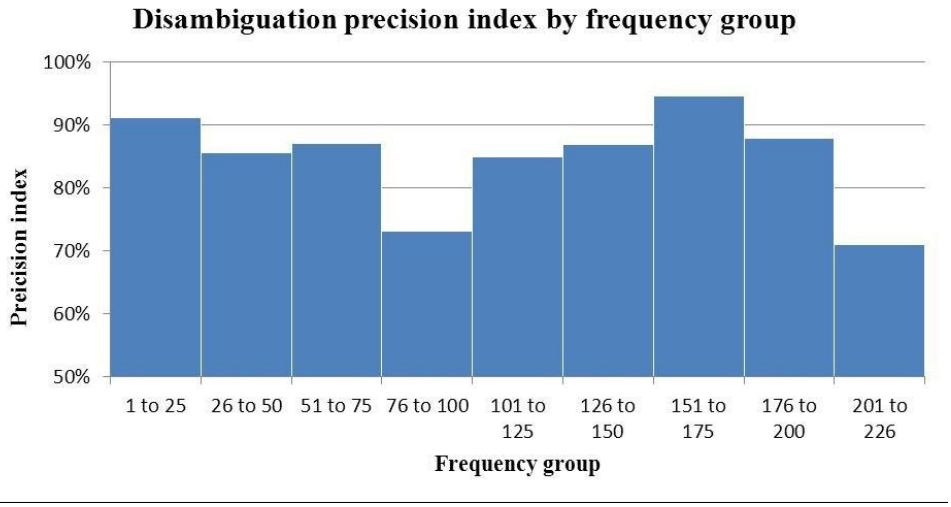

Figure 1. Disambiguation precision index by frequency group

As can be noted, the rate of correct classification of HH pairs tends to stay above 85% (only two frequency groups "76 to 100" and "201 to 226", were below 85 %). Two frequency groups were above 90% and the average, as already mentioned, was 90.1%. This discrepancy is justified by the data distribution, although only two out of nine frequency groups were above 90%, the first frequency group plays a primary role in HH disambiguation, containing 14,061 of the 16,851 tokens analyzed. MXPOST proves to be a reliable tool for classifying grammatical classes of HH pairs. Using MXPOST in conjunction with the above-mentioned two rules, it is possible to obtain good performance in HH disambiguation (at least in the case of mid vowels found in stressed syllables). The corpus used shows that cases of these pairs correspond to approximately 1.44% of all occurrences within the corpus. Such an occurrence rate, although relatively low, shows to be relevant, seeing that if their grammatical classes are not considered, these words could be incorrectly transcribed; thus, severely damaging the results of applications employing speech processing technologies.

The forty most frequent HH pairs were analyzed as well, individually, as to the degree of belonging to each grammar class. In other words, we verified the distribution of each individual HH pair regarding their parts of speech. The chart below shows the ratio values of each grammatical class derived from the forty pairs.

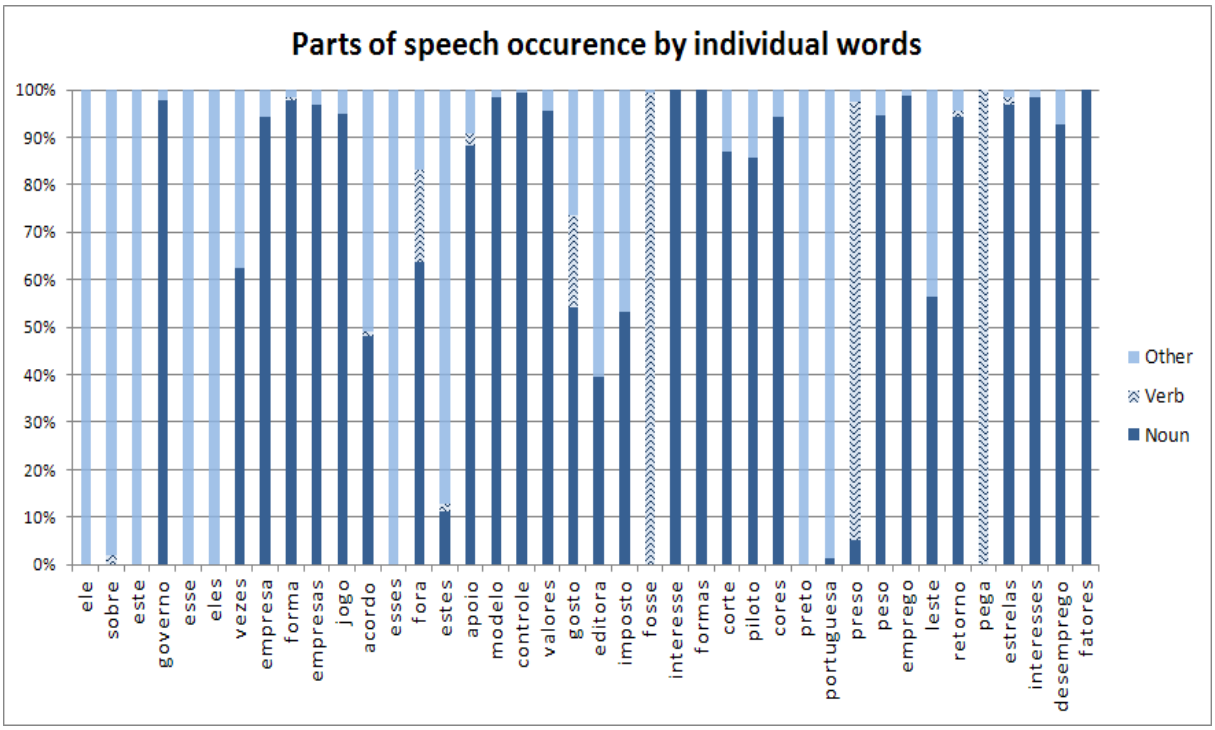

Figure 2: Parts of speech occurrence by individual words

As noted, there are few words that have grammatical category ratios close to 1:1. In most pairs, a predominance trend of one grammatical class exists, in this case, the non-verb class. There are several pairs with predominance of the non-verb classes exceeding 95% of all occurrences. The data therefore demonstrates that, in many cases of homography and heterophony, noun and pronoun classes take the majority. This fact can be exploited in speech recognition, in order to increase transcription accuracy: in dubious cases, where there is uncertainty as to which class an HH belongs, it is more likely to be correct if the non-verb class attribute is awarded.

However, one other plausible justification for the results obtained is related to the type of corpus used. The MAC-Morpho is a corpus consisting of newspaper articles. In light of this, it is known that journalism genres rarely make use of verbs in the 1st person singular, since the use of the first person in a journalistic text entails subjectivity. Tavares (1997), investigating verb usage in the genres of newspaper and TV reporting, noted that the occurrence of the 1st person singular was only 2% for the first genre studied, and 0% for the second. Tavares (1997) notes that, in journalism texts, the 3rd person singular and 3rd person plural dominated, accounting for 64% and 33% respectively in newspapers and 89% and 11% in TV reports. Therefore, the prevalence of non-verb classes noticed in our data may have been led to a certain degree by the nature of the corpus used. Many of the HH pairs we observed involve precisely the alternation between the 1st person singular of a verb and another non-verb class. It should be mentioned; therefore, that the low occurrence of first person, singular verbs in journalism genres, can have possible bias against our data due

to the use of a newspaper corpus. It would be worthwhile; therefore, to replicate this study, taking many HH pairs into account like: “apoio”, “jogo”, “acerto”, which have open vowels as verbs and are likely to occur with greater frequency in other text genres. (e.g., <eu apoio>, <eu jogo>, <eu acerto>). A similar call can be made in general for more work on freely available corpora of BP.

4. Final considerations

The study of HH pairs involving openness in mid vowels in BP is of interest to researchers developing tools for this language. In light of the issues presented in high quality systems like Veiga et al. (2011), where over 80% of grapheme to phone(me) conversion errors were caused by mid vowel shifts in stressed syllables, it is certainly an area which could be greatly improved. The results obtained in this paper can help develop various tasks such as, automatic processing of speech and CALL tools for teaching pronunciation. So far, most of the conversion systems employed in text-to-speech applications for BP have not devoted sufficient attention to this issue, but simply present considerations in passing about the problems that could and/or do arise in their systems with respect to the proper pronunciation of <e> <o> stressed syllable vowels. The suggestion to create exception dictionaries shows up as a desirable solution; however, it is ad hoc and not expandable beyond words not registered in the dictionary. Thus, this work contributes to the research area, as it shows that the majority of cases involving HH pairs containing mid vowels can be solved automatically with just two rules, when information about the word class is taken into account. In this study, the tagger, MXPOST, was used to perform the automatic part of speech tagging, achieving a high degree of accuracy (90.1%), thus proving itself a useful tool for the disambiguation of HH pairs; although, in future studies, we plan to investigate other tagger options in hopes of still better results. The results show a trend that in many cases of homographic heterophony, there is a certain level of non-verb class majority, with various pairs showing a non-verb class predominance of over 95%. However, this trend may require further research, specifically investigating the possibility that the present corpus bias, regarding the use of the 1st person singular, is due to its journalistic nature.

In summary, this paper proposed a disambiguation solution for the majority of HH mid vowel pairs, seeking to offer, especially in the areas of speech synthesis and recognition, a contribution towards solving this often cited issue in systems for BP.

As for future work, the task of classifying HH pairs with stressed mid vowel changes does not stop here. The authors of this paper plan to investigate other methods to account for as many pairs as possible. Nearly thirty linguistic based rules have already been established, which cover all word pairs belonging to this family. These rules have been formalized so that concrete applications can be developed to treat the errors; hopefully able to effectively treat nearly all pairs. An example of such an application is a feature extractor using line spectral frequencies which is a technology commonly used in electrical signal processing and is often present in applications commercially used by telephone companies to recognize spoken digits [Silva et al. 2013]. This type of classifier could possibly be

adapted to phonemic feature extraction by applying a technique that allows for very small windows of data to be extracted and analyzed, even when their signals are very similar. These windows would be mapped to phonological rules of BP in instances of open and closed vowels, in order to disambiguate them. Such classifiers have achieved accuracy rates exceeding 89% for BP in spoken digit recognition [Silva et al. 2013].

References

Aires, R. V. X., Aluísio, S. M., Kuhn, D. C. S., Andreeta, M. L. B. and Oliveira Jr., O. N. (2000) “Combining Multiple Classifiers to Improve Part of Speech Tagging: A Case Study for Brazilian Portuguese”. In the Proceedings of the Brazilian AI Symposium (SBIA’2000).

Aluísio, S. M., Pelizzoni, J. M., Marchi, A. R., Oliveira, L. H., Manenti, R. and Marquiafável, V. (2003). “An Account Of The Challenge Of Tagging A Reference Corpus Of Brazilian Portuguese”. In: Propor´2003, 2003, Faro. Lecture Notes On Artificial Intelligence. Proceedings Of Propor´2003. Springer Verlag, v. 1.

Bick, E. (2000). “The Parsing System Palavras - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework”. Aarhus: Aarhus University Press.

Bortoni-Ricardo S. M., Malvar, E. and Gomes, C. A. (1992). “A variação das vogais médias pretônicas no português do Brasil: um fenômeno neogramático ou de difusão lexical”. In: Revista de Estudos da Linguagem, Faculdade de Letras - UFMG, v. 1, pp. 9-30.

Braga, D. (2008). “Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto Fala em Português”. PhD Thesis. University of A Coruña, A Coruña, Spain. (Directed by University of Coruña, co-directed by University of Minho and Federal University of Rio de Janeiro).

Braga, D. and Marques, M. A. (2007). “Desambiguação de homógrafos para Sistemas de conversão Texto-Fala em Português”. In: Diacrítica, 21.1 (Série Ciências da Linguagem) Braga: CEHUM/Universidade do Minho, pp 25-50.

Cristófaro-Silva, T. and Almeida, L. S. (2005). ASPA: “A formulação de um banco de dados de referência da estrutura sonora do português contemporâneo”. In: Anais do XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo: Sociedade Brasileira de Computação, 2005. v. 1. p. 2268-2277.

Ferrari, L., Barbosa, F. and Resende Jr., F. G. V. (2003) "Construções gramaticais e sistemas de conversão texto-fala: o caso dos homógrafos”. In: Proceedings of the International Conference on Cognitive Linguistics, v. 1.

Frunza, O. and Inkpen, D. (2009) Identification and Disambiguation of Cognates, False Friends, and Partial Cognates Using Machine Learning Techniques, International Journal of Linguistics, vol. 1, no. 1, p. 1-37, Otawa, Canada.

Miranda, A. R. (2006). “Um estudo sobre a aquisição ortográfica das vogais do português”. Anais da ANPEDSul – UFSM, Santa Maria.

Núcleo Interinstitucional de Linguística Computacional - NILC. (2000) NILC's Taggers. http://www.nilc.icmc.usp.br/nilc/tools/nilctaggers.html. Apr. 23, 2013.

Seara, I., Kafka, S., Klein, S. and Seara, R. (2001). “Considerações sobre os problemas de alternância vocálica das formas verbais do Português falado no Brasil para aplicação em um sistema de conversão Texto-Fala”, SBrT 2001 – XIX. Simpósio Brasileiro de Telecomunicações. Fortaleza, Brasil.

Seara, I., Kafka, S., Klein, S. and Seara, R. (2002). “Alternância vocálica das formas verbais e nominais do Português Brasileiro para aplicação em conversão Texto-Fala”, Revista da Sociedade Brasileira de Telecomunicações, vol. 17, n.1, pp. 79- 85.

Silva, D., Alves de Souza, V. and Batista, G. (Forthcoming). “A comparative study between MFCC and LSF coefficients in automatic recognition of isolated digits pronounced in Portuguese and English”.

Silva, D. C.; Braga, D.; Resende, F.G.V. (2009) “Conjunto de regras para desambiguação de homógrafos heterófonos do português brasileiro”. In: Simpósio Brasileiro de Telecomunicações, 2009, Blumenau. XXVII Simpósio Brasileiro de Telecomunicações, v. 1. p. 1-6.

Silva, D. C. (2011) “Algoritmos de processamento da linguagem e síntese de voz com emoções aplicados a um conversor texto-fala baseado em HMM”. Tese de Doutorado. Programa de pós-graduação em Engenharia Elétrica, Universidade Federal do Rio de Janeiro, RJ.

Silva, P., Batista, P., Neto, N. and Klautau, A. (2010) “An Open-Source Speech Recognizer For Brazilian Portuguese With A Windows Programming Interface”. In: The International Conference On Computational Processing Of Portuguese , Porto Alegre.

Ratnaparkhi, A. (1996) “A Maximum Entropy Part-Of-Speech Tagger”. In the Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania.

Robinson, A. (2006). The story of writing. London: Thames and Hudson. pp. 224.

Roces. L. (2010). “Efeitos de Coprodução V_V sobre os Deslocamentos das Vogais Médias no Espaço Vocálico do Português Brasileiro”. Tese (Doutorado em Linguística) - Universidade Estadual de Campinas, pp. 100.

Tavares, M. A. (1997) "O verbo no texto: notícia e reportagem". In: Working papers em Linguística. n. 1. jul/dez. Florianópolis: UFSC.

Tomaz, K. (2006). “Alternância de vogais médias posteriores em formas nominais de plural no português de Belo Horizonte Revisitando a metafonia nominal”. Dissertação (Mestrado em Letras) - Universidade Federal de Minas Gerais, pp. 165.

Viegas, M. C. (2002). “O alçamento de vogais médias pretônicas e o conceito de léxico com armazenamento exemplar”. In: I Congresso Internacional de Fonética e Fonologia. Caderno de Resumos do I Congresso Internacional de Fonética e Fonologia. Belo Horizonte: Faculdade de Letras da UFMG.

Veiga, A., Candeias, S. and Perdigão, F. (2011) "Conversão de Grafemas para Fonemas em Português Europeu – Abordagem HÍbrida com Modelos Probabilísticos e Regras Fonológicas". J. J. Almeida, A. Simões, X. Guinovart (eds.). In: LinguaMÁTICA, Revista para o Processamento Automático das Línguas Ibéricas, vol 3, nº2: 39-51.