Conjunto de Regras para Desambiguação de Homógrafos Heterófonos no Português Brasileiro

Denilson C. Silva, Daniela Braga e Fernando G. V. Resende Jr.

***Resumo*— Este artigo apresenta um conjunto de regras para decidir a leitura de homógrafos heterófonos aplicados em um sistema TTS do português brasileiro. O método proposto é composto de uma análise morfossintática, que trata problemas com homógrafos pertencentes a diferentes classes gramaticais e uma análise semântica, que trata problemas com homógrafos de mesma classe gramatical, através de 107 pares de homógrafos organizados em 22 algoritmos. Os algoritmos foram implementados e testados com textos de três naturezas, apresentando taxas de erro de 1,42%, 1,40% e 1,30%, respectivamente.**

***Palavras-Chave*— Texto-Fala, Homógrafo, Síntese de Voz, Análise Morfossintática, Análise Semântica**

***Abstract*— This paper presents a rule-based algorithm set used to decide the reading of homographs applied in a Brazilian Portuguese Text-to-Speech (TTS) system. The proposed approach is composed of a morphosyntactic analysis, which deals with problems of homographs that belong to different part-of-speech (POS), and a semantic analysis, which deals with problems of homographs that belong to the same POS, through 107 homograph pairs, which are organized into 22 disambiguation algorithms. The algorithms were implemented and tested with three types of texts. Computer experiments show they have achieved an error rate of 1.42%, 1.40% and 1.30%, respectively.**

***Keywords*— Text-to-Speech, Homograph, Speech Synthesis, Morphosyntactic Analysis, Semantic Analysis**

I. INTRODUÇÃO

A distinção entre homógrafos heterófonos é um problema complexo na transcrição grafema-fone, porque a saída da transcrição fonética não é única para cada homógrafo, pois o algoritmo que faz a transcrição precisa decidir entre duas situações possíveis: ou a vogal tônica é aberta ([E]/[O]) ou ela é fechada ([e]/[o]) [1]. Palavras como <seca> (substantivo, "a s[e]ca", e verbo, "ele s[E]ca",) e <sede> (substantivo, "a s[e]de", e também substantivo, "a s[E]de"), possuem a mesma grafia, porém significados diferentes. Por este motivo, elas são chamadas de homógrafos heterófonos e se não forem analisadas corretamente, podem comprometer a transcrição fonética e, consequentemente, a qualidade da síntese. Diminuir os erros na leitura de homógrafos, melhora significativamente a qualidade do sintetizador, bem como sua naturalidade e inteligibilidade.

Em geral, podemos ter homógrafos que pertencem a diferentes classes gramaticais (POS - *part-of-speech*) (<seca>) e homógrafos pertencentes a mesma classe gramatical (<sede>), conforme exemplificado anteriormente. Desta forma, para uma conversão grafema-fone adequada, o uso de regras linguísticas para coleta de informações morfossintáticas, no caso de homógrafos de diferentes classes gramaticais, e de informações semânticas, para o caso de homógrafos com mesma classe, são um bom recurso para decidir se a palavra possui vogal tônica aberta ou fechada.

A desambiguação de homógrafos é um assunto bastante explorado pela comunidade científica dada a sua importância em síntese de voz em diversos idiomas. Em [2] temos uma tipologia de pares de homógrafos na língua inglesa e algumas técnicas para desambiguação tradicionalmente utilizadas, tais como *n-gram*, *taggers*, classificadores *bayesianos* e árvore de decisão, bem como a proposta de um sistema híbrido, combinando as três melhores técnicas descritas. Em [3], o assunto é tratado em idiomas como tailandês, chinês e japonês, onde as palavras não possuem delimitação de fronteira. Uma técnica de reconhecimento de padrões, denominada *winnow*, é utilizada para fazer a segmentação das palavras e resolver o problema de ambiguidade dos homógrafos. Em [4], os autores apresentam um estudo da relação entre caracteres chineses e sua pronúncia, bem como propõem uma solução para a desambiguação de caracteres polifônicos. Em relação à desambiguação de homógrafos em sistemas de conversão texto-fala no português europeu, nas referências [5] e [6] são analisadas as melhorias que podem ocorrer no desempenho da conversão texto-fala, através da influência de informações morfossintáticas na desambiguação de homógrafos. Em [7] a desambiguação aborda também a coleta de informações semânticas no português europeu, além de comparar a metodologia com redes neuronais. Para o português brasileiro, em [8], [9] um analisador morfossintático é apresentado para solucionar o problema de alternâncias vocálicas entre substantivos e verbos, sem, no entanto, abordar a desambiguação de homógrafos semanticamente. Já em [10], [11] e [12], temos ambas as abordagens, morfossintática e semântica, mas o método foi testado apenas com um único exemplo.

Neste trabalho, apresentamos um método de desambiguação de homógrafos heterófonos aplicado a um sistema de conversão texto-fala para o português brasileiro [13], que soluciona a ambiguidade de um número amplo de pares de homógrafos, tanto morfossintaticamente como semanticamente [14]. O algoritmo proposto foi implementado e testado com segmentos de texto extraídos aleatoriamente de base de dados de texto com naturezas distintas, a saber: CETEN-Folha (natureza jornalística) [15], Bíblia Cristã (natureza formal e religiosa) [16] e da obra literária "Dom Casmurro", de

Denilson C. Silva¸ Programa de Engenharia Elétrica, PEE/COPPE, UFRJ, Rio de Janeiro, RJ, E-mail: denilson@lps.ufrj.br

Daniela Braga, Microsoft Language Development Center, Porto Salvo, Portugal, E-mail: i-dbraga@microsoft.com

Fernando G. V. Resende Jr, DEL/Escola Politécnica, PEE/COPPE, UFRJ, Rio de Janeiro, RJ, E-mail: gil@lps.ufrj.br

Machado de Assis (natureza histórico-literária) [17]. Os resultados alcançaram taxas de erro de $1,42\%$, $1,40\%$ e $1,30\%$, respectivamente.

Este artigo está organizado da seguinte forma: a Seção II apresenta a metodologia aplicada na desambiguação de homógrafos. Na Seção III, apresentamos as bibliotecas utilizadas nos algoritmos de desambiguação. Na Seção IV, temos alguns algoritmos baseados nas regras linguísticas utilizadas. Na Seção V, descrevemos os testes realizados, bem como a base de dados empregada. Na Seção VI são apresentadas as conclusões do artigo e trabalhos futuros.

II. Metodologia aplicada na desambiguação

O algoritmo proposto para desambiguação de homógrafos é apresentado na Figura 1. Este algoritmo é, em essência, um conjunto de regras linguísticas, separadas por tipo, baseado em um método composto de duas partes: a análise morfossintática, que soluciona problemas com homógrafos de classes gramaticais diferentes e a análise semântica, que soluciona problemas com homógrafos de mesma classe.

A metodologia aplicada neste trabalho partiu da coleta de pares de homógrafos existentes no português brasileiro, resultando num conjunto de 107 pares de homógrafos, que pode ainda ser atualizado, já que temos a versatilidade de acrescentar quantos homógrafos forem necessários na biblioteca de homógrafos existente. O conjunto de homógrafos está organizado baseado em regras por tipos, de acordo com a oposição gramatical existente e a alternância fonética no par de homógrafos. Para cada tipo de homógrafo, temos um algoritmo com várias perguntas relativas ao contexto morfossintático do homógrafo, que vão contribuir para a correta transcrição fonética da vogal tônica.

Para a realização das perguntas nos algoritmos e a determinação das classes gramaticais das palavras vizinhas ao homógrafo, várias bibliotecas foram coletadas, formando um analisador morfológico.

Na Tabela I, temos o conjunto de homógrafos usados neste trabalho, separados por tipo.

III. Bibliotecas empregadas na análise do texto

Várias bibliotecas foram usadas na análise morfossintática e na análise semântica:

1) Biblioteca de homógrafos, contendo 107 pares de homógrafos agrupados em 22 tipos, que podem ser vistos na Tabela I.
2) Biblioteca de Classes Fechadas, contendo as classes gramaticais que têm um número fixo de componentes (pronomes, preposições, advérbios, conjunções, contrações, artigos, numerais, interjeições).
3) Biblioteca de Morfemas, contendo sufixos nominais, verbais, adjetivais e adverbiais, prefixos e radicais gregos e latinos.
4) Biblioteca de Lemas, contendo o dicionário em português *Jspell* com aproximadamente 34000 palavras morfologicamente anotadas [18].
5) Biblioteca de verbos irregulares, contendo a flexão dos principais verbos irregulares existentes no português brasileiro.

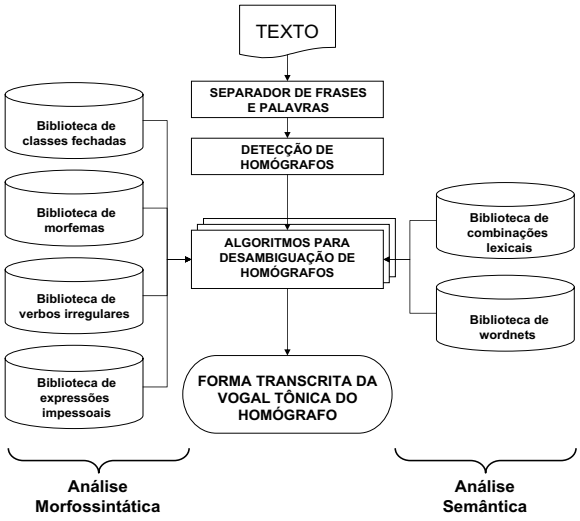

Fig. 1. Estrutura do desambiguador de homógrafos [14].

TABELA I

Conjunto de homógrafos separados por tipo.

Tipo	Homógrafo
1	acerto, apelo, aperto, apreço, começo, concerto, conserto, desemprego, desespero, emprego, enredo, erro, esmero, espeto, flagelo, gelo, governo, interesse, interesses, modelo, pego, peso, rego, selo, testo e zelo.
2	aborto, acordo, adorno, aforro, almoço, arrojo, arroto, choco, choro, conforto, consolo, contorno, controle, coro, desgosto, despojo, destroço, encosto, endosso, esforço, estorvo, folgo, gosto, jogo, logro, namoro, olho, piloto, reforço, rodo, rogo, rolo, sopro, suborno, sufoco, toco, toldo, topo, torno, troco e troço.
3	rola e rolha.
4	colher e meta.
5	desses, deste e destes.
6	fora.
7	seco, seca e secas.
8	boto.
9	este.
10	leste.
11	sobre.
12	rota, rotas, tola e tolas.
13	corte, cortes, forma, formas, molho e soco.
14	cerca.
15	pega e pegas.
16	besta e bestas.
17	sede e sedes.
18	medo e medos.
19	termos.
20	cor.
21	lobo e lobos.
22	bola e bolas.

6) Biblioteca de expressões impessoais, contendo expressões com o verbo ser na terceira pessoa + adjetivo.
7) Biblioteca de combinações lexicais restritas, contendo expressões idiomáticas, provérbios ou expressões fixas de uma ou mais palavras. Esta biblioteca é usada na análise semântica.
8) Biblioteca de *Wordnets*, desenvolvidas sob o conceito de *Wordnet* [19], contendo palavras que são semanticamente e cognitivamente relacionadas com o homógrafo analisado. Esta biblioteca também é usada na análise semântica.

O texto é inicialmente dividido em palavras e frases. Em seguida, o sistema realiza a busca por homógrafos, através da biblioteca existente. Uma vez identificado um homógrafo, este é conduzido ao algoritmo correspondente ao seu tipo.

Na Tabela II e na Tabela III, podemos ver homógrafos pertencentes a diferentes classes gramaticais, bem como pertencentes à mesma classe gramatical, respectivamente. Como as tabelas mostram, as oposições gramaticais que mais ocorrem são entre substantivo e verbo, do ponto de vista morfológico, e entre [e]/[E] e [o]/[O], do ponto de vista fonético. Uma evidência sistemática é que em substantivos, a vogal tônica é tipicamente fechada, embora nas formas verbais a vogal tônica seja aberta. Homógrafos do Tipo 1 e 2 representam $62,6\%$ do total de homógrafos da biblioteca. Homógrafos do Tipo 13, 14, 15 e 19 necessitam tanto da análise morfológica como da análise semântica, uma vez que eles podem desempenhar três possibilidades na saída.

TABELA II

EXEMPLOS DE HOMÓGRAFOS COM DIFERENTES CLASSES GRAMATICAIS.

Tipo	Alternância vocálica e Oposição gramatical	Exemplo
1	[e] Substantivo / [E] Verbo	Nosso erro foi muito grande. Eu erro bastante.
2	[o] Substantivo / [O] Verbo	Ele fechou o olho esquerdo. Eu olho para cima.
3	[o] Substantivo / [O] Verbo	Eu vi uma rola branca. Ele deita e rola.
4	[e] Substantivo / [E] Verbo	É época de colher o tomate. Essa é a nossa meta.
5	[e] Contração / [E] Verbo	Ele ganhou dois desses prêmios. Era bom que nunca desses a notícia.
6	[o] Verbo / [O] Advérbio	Ele fora uma pessoa honesta. Eu estou fora do jogo.
7	[e] Adjetivo ou Substantivo / [E] Verbo	O rio estava muito seco. Eu seco os pés na entrada.
8	[o] Adjetivo ou Substantivo / [O] Verbo	Ele viu um boto na praia. Eu boto azeite na salada.
9	[e] Demonstrativo / [E] Adjetivo ou Substantivo	Este armário é meu. Norte, sul, este, oeste.
10	[e] Verbo / [E] Adjetivo ou Substantivo	Leste a notícia? Seguiu para o leste.
11	[o] Preposição / [O] Verbo	Comentou sobre o fato. É bom que sobre uma garrafa.
12	[o] Adjetivo ou Verbo [O] Substantivo /	Ela andava toda rota. Nós seguimos a rota.
13	[o] Substantivo / [O] Verbo / Substantivo	Ela comprou pão de forma. De qualquer forma iremos ao passeio.
14	[e] Preposição / Substantivo / [E] Verbo	Eles andaram cerca de dez kilômetros. Ele cerca seu terreno com arame farpado.
15	[e] Substantivo / [E] Verbo / Substantivo	Aquela ave parece uma pega. Olha que essa moda ainda pega.

TABELA III

EXEMPLOS DE HOMÓGRAFOS COM MESMA CLASSE GRAMATICAL.

Tipo	Alternância vocálica e Oposição gramatical	Exemplo
16	[e] Substantivo / [E] Substantivo	Ele é metido a besta. Ele conseguia disparar a besta.
17	[e] Substantivo / [E] Substantivo	Ele estava com uma sede insuportável. A sede da empresa fica em Paris.
18	[e] Substantivo / [E] Substantivo	Ela estava com medo de morrer. Eles venceram todo o Império Medo-Persa.
19	[e] Substantivo / Verbo [E] Substantivo	Estes são os nossos termos. A termos tinha café quente.
20	[o] Substantivo / [O] Substantivo	O vestido era cor de rosa. Sabia tudo de cor e salteado.
21	[o] Substantivo / [O] Substantivo	Na estória não tinha lobo mau. Ele feriu o lobo temporal.
22	[o] Substantivo / [O] Substantivo	Só amassei a bola de carne. Eu não tenho bola de cristal.

IV. APLICAÇÃO DAS REGRAS PROPOSTAS

Depois das sentenças serem divididas em palavras e frases, o sistema busca candidatas a homógrafos e as relaciona com sua biblioteca de homógrafos. Se o sistema identifica uma palavra como sendo homógrafo, ele a relaciona com um tipo. Cada tipo possui um algoritmo correspondente com perguntas sobre o contexto de um determinado homógrafo. As bibliotecas existentes são utilizadas para esta tarefa. Nas Tabelas IV e V apresentamos, respectivamente, os pseudo-códigos que realizam a desambiguação dos homógrafos "cerca" e "sede/sedes". Os símbolos usados nos algoritmos podem ser vistos na Tabela VI. O primeiro conjunto de perguntas conduz à saída mais provável. Se a resposta for negativa, o sistema realiza outra seção de perguntas e, caso a resposta seja positiva, a saída é a ocorrência estatisticamente menos provável. Se ainda for negativa, o sistema conduz ao caso padrão (default).

V. TESTES REALIZADOS COM O CONJUNTO DE REGRAS

A. Base de dados utilizada

O sistema foi testado com extratos de texto provenientes de três bases com naturezas distintas (jornalística, formal-religiosa e histórico-literária). Estas vertentes foram selecionadas visando explorar o rigor gramatical, além de ter a possibilidade de encontrar alguns homógrafos e testar o respectivo algoritmo, já que seria necessário uma quantidade enorme de palavras para poder encontrar todos num mesmo texto. A seguir são definidas as origens de cada fonte de texto:

1) CETEN-Folha - Esta base de dados é um corpus contendo cerca de 24 milhões de palavras do português brasileiro, criado pelo projeto Processamento Computacional do Português com base nos textos do jornal Folha de São Paulo e que fazem parte do corpus NILC/São Carlos, compilado pelo Núcleo Interinstitucional de Linguística Computacional (NILC) [15]. O teste realizado com esta base foi composto de um segmento de texto contendo 1.564.591 palavras, onde foram identificadas 13.246 homógrafos, ou seja, $0,85\%$ do texto são homógrafos.
2) Bíblia Cristã - Esta base de dados é uma versão em formato texto da Bíblia Sagrada para o português brasileiro [16], composta de mais de 750.000 palavras, apresentando um estilo lingüístico mais elaborado do que a base do CETEN-Folha, motivo pelo qual foi incluída no teste dos algoritmos. Além disso, pelo estilo existente, alguns homógrafos puderam ser encontrados em maior número e, portanto, ter os seus algoritmos mais solicitados. O teste, neste caso, foi realizado com toda a base e 5.558 homógrafos foram identificados ($0,74\%$ do texto processado).
3) Dom Casmurro - Esta obra literária é um romance do

TABELA IV

PSEUDO-CÓDIGO DO ALGORITMO DE DESAMBIGUAÇÃO DE HOMÓGRAFOS DO TIPO 14.

```
1: if (A palavra é homógrafo do tipo 14) then
2:    if (Homógrafo pertence a BC_cerca_e) ou (WN_cerca_e existe em F0) ou (P+2 ou P+3 = NUM) then
3:       V = [e]
4:    else if (P-1 = <uma>, <a>, CONTR ou PREPO) ou (P+2 = <madeira>, <arame>, <espinhos>) ou (<saltar> ou <levantar>
      existe em F0) ou (P+1 = ad) then
5:       V = [e]
6:    else if (P-1 ou P-2 = <que>, <não>, <ainda>, <já> ou <também>) ou (P-1 = <ele>, <ela> ou P_PESS_O_1) then
7:       V = [E]
8:    else
9:       V = [e]
10:   end if
11: else
12:    Vá para o algoritmo 15
13: end if
```

TABELA V

PSEUDO-CÓDIGO DO ALGORITMO DE DESAMBIGUAÇÃO DE HOMÓGRAFOS DO TIPO 17.

```
1: if (A palavra é homógrafo do tipo 17) then
2:    if (WN_sede_e existe em F-1, F0 ou F+1) ou (Homógrafo pertence a BC_sede_e) then
3:       V = [e]
4:    else if (WN_sede_E existe em F-1, F0 ou F+1) ou (Homógrafo pertence a BC_sede_E) then
5:       V = [E]
6:    else
7:       V = [E]
8:    end if
9: else
10:    Vá para o algoritmo 18
11: end if
```

TABELA VI

SIMBOLOGIA USADA NOS ALGORITMOS DE DESAMBIGUAÇÃO.

Símbolo	Significado
P-1, P-2	Última e penúltima palavra, respectivamente.
P+1	próxima palavra.
F0, F-1, F+1	frase atual, última frase e próxima frase, respectivamente.
PREPO	preposição.
CONTR	contração.
P_PESS_O_1	pronome pessoal oblíquo (<me>,<mim>,<te>, <ti>,<se>,<si>,<nos>, <vos>,<lhe(s)>, <no-lo(s)>,<no-la(s)>, <vo-lo(s)>, <vo-la(s)>, <lho(s)> ou <lha(s)>).
NUM	numeral.
BC	combinações lexicais restritas.
WN	wordnet.
V	vogal tônica do homógrafo.

escritor brasileiro Machado de Assis e serviu como base de dados de texto contendo quase 70.000 palavras. Foi publicado em 1899 e a história se passa no Rio de Janeiro do Segundo Império [17]. É um romance psicológico, narrado em primeira pessoa por Bentinho, permitindo assim encontrar alguns dos homógrafos dos tipos 1 e 2, que são verbos, para possibilitar o teste aos algoritmos. O fato do texto ser histórico e literário nos leva a ter um conteúdo com uma linguística bem elaborada, servindo como um bom teste para os algoritmos de desambiguação. Por fim, neste teste foi identificado um total de 385 homógrafos, ou seja, $0,55\%$ do texto total da base.

B. Resultados obtidos

Conforme os textos apresentados na Subseção V-A, foram realizados testes com as regras de desambiguação de homógrafos. Os resultados podem ser verificados nas Tabelas VII, VIII e IX. Podemos verificar que, apesar da variação no tamanho dos textos submetidos ao teste e do número de homógrafos encontrados, os resultados mostraram uma regularidade no percentual de acerto global do algoritmo proposto. É verificado também que apenas no teste realizado com o conteúdo de natureza jornalística, foram encontrados homógrafos pertencentes a todos os tipos listados neste trabalho.

Considerando os resultados obtidos, podemos notar que o resultado geral apresenta uma taxa de erro com valores consideravelmente pequenos (cerca de $1,37\%$ em média), porém não podemos deixar de lado o fato de que ainda existem alguns homógrafos com ocorrências muito pequenas que acabam ponderando os seus erros individuais para valores muito elevados. Isto é um fator que demonstra a complexidade de solução do problema dos homógrafos.

Os erros ocorridos nos algoritmos dos tipos 1 e 2 apresentaram um índice elevado de erros quando seguidos, principalmente, por preposições ou contrações, antecedidos por formas verbais flexionadas.

Nos demais casos, podemos notar que os índices de erro foram muito pequenos ou não existiram. Os erros que ainda assim ocorreram, podem ser justificados por algumas falhas ortográficas existentes na base de dados ou pelo fato de

TABELA VII

TESTES COM O ALGORITMO PROPOSTO - CETEN-FOLHA.

Tipo	Ocorrência	Número de erros	% de erro
1	3409	44	1,29%
2	2640	69	2,61%
3	11	1	9,10%
4	95	5	5,26%
5	637	1	0,16%
6	482	11	2,28%
7	90	10	11,11%
8	5	0	0,00%
9	825	0	0,00%
10	169	0	0,00%
11	2335	14	0,60%
12	47	2	4,26%
13	826	13	1,57%
14	866	3	0,35%
15	43	0	0,00%
16	11	1	9,10%
17	148	7	4,73%
18	130	0	0,00%
19	108	7	6,48%
20	68	0	0,00%
21	39	0	0,00%
22	262	0	0,00%
TOTAL	13246	188	**1.42%**

TABELA VIII

TESTES COM O ALGORITMO PROPOSTO - BÍBLIA CRISTÃ.

Tipo	Ocorrência	Número de erros	% de erro
1	209	4	1,91%
2	321	11	3,42%
3	5	1	20,00%
4	27	2	7,41%
5	333	12	3,60%
6	428	6	1,40%
7	61	5	8,20%
8	0	—	—
9	984	0	0,00%
10	5	1	20,00%
11	2740	14	0,51%
12	11	1	9,10%
13	65	4	6,15%
14	51	2	3,92%
15	5	0	0,00%
16	46	1	2,17%
17	107	10	10,60%
18	82	1	1,22%
19	60	2	3,33%
20	3	1	33,30%
21	14	0	0,00%
22	1	0	0,00%
TOTAL	5558	78	**1,40%**

existirem frases isoladas ao longo do texto, dificultando uma avaliação contextual mais precisa. Grande parte do sucesso obtido ocorre devido ao fato de que se todas as respostas falharem, o sistema sai com a transcrição padrão. Apesar de o percentual de ocorrência de homógrafos, em geral, ser inferior a $1,0\%$, para o objetivo deste trabalho, que é a síntese de voz, basta que uma palavra seja transcrita de forma equivocada, para que o sistema não soe agradável ao ouvinte. As naturezas das bases de dados também explicam porque alguns dos homógrafos foram muito pouco encontrados, como é o caso de b[o]to / b[O]to (algoritmo 8), comum no português brasileiro, porém muito informal ou específico.

VI. CONCLUSÕES

Neste trabalho, apresentamos um conjunto de algoritmos baseados em regras linguísticas, separados por tipo, para desambiguação de homógrafos heterófonos aplicados à conversão texto-fala no português brasileiro. O método proposto é capaz de realizar a distinção entre 107 pares de homógrafos existentes no português brasileiro, organizados em 22 tipos de algoritmos de desambiguação, com a possibilidade de ter sua performance aumentada, à medida que novos homógrafos são inseridos na base de dados de homógrafos. O conjunto de algoritmos proposto foi implementado e testado com extratos de textos com naturezas distintas, atingindo uma taxa de erro de $1,42\%$ para o de natureza jornalística, $1,40\%$ para o de natureza formal-religiosa e $1,30\%$ no conteúdo com natureza histórico-literária. Estes resultados são bastante animadores, considerando a diversificação da natureza dos textos utilizados nos testes. Os algoritmos propostos podem resolver a ambiguidade de vários pares de homógrafos existentes no português brasileiro. Como trabalhos futuros, desejamos conduzir os testes com extratos de textos ainda maiores e de diferentes naturezas, avaliar a necessidade da inclusão de homógrafos, bem como comparar este sistema com técnicas estatísticas *data-driven*.

TABELA IX

TESTES COM O ALGORITMO PROPOSTO - DOM CASMURRO.

Tipo	Ocorrência	Número de erros	% de erro
1	36	0	0,00%
2	71	1	1,41%
3	0	—	—
4	3	0	0,00%
5	30	0	0,00%
6	52	2	3,85%
7	6	0	0,00%
8	0	—	—
9	86	0	0,00%
10	0	—	—
11	35	0	0,00%
12	2	1	50,00%
13	5	0	0,00%
14	2	0	0,00%
15	1	0	0,00%
16	7	0	0,00%
17	1	0	0,00%
18	22	0	0,00%
19	5	1	20,00%
20	17	0	0,00%
21	0	—	—
22	4	0	0,00%
TOTAL	385	5	**1,30%**

REFERÊNCIAS

[1] *Speech Assessment Methods Phonetic Alphabet*, Speech Assessment Methods Phonetic Alphabet (SAMPA), http://www.phon.ucl.ac.uk/home/sampa/home.htm, acessado em 23/02/2008.

[2] D. Yarowsky, *Homograph disambiguation in text-to-speech synthesis*, In Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors, Progress in Speech Synthesis, pp. 157-172, 1997.

[3] V. Tesprasit and P. Charoenpornsawat and V. Sornlertlamvanich, *A Context-Sensitive Homograph Disambiguation in Thai Text-to-Speech Synthesis*, In Proc. of HLT-NAACL'03 - Human Language Technology and North American Chapter of the Association for Computational Linguistics Conference, 2003.

[4] H. Dong and J. Tao and B. Xu, *Grapheme-to-phoneme conversion in Chinese TTS system*, International Symposium on Chinese Spoken Language Processing, pp. 165-168, 2004.

[5] R. Ribeiro and L. C. Oliveira and I. Trancoso, *Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese*, PROPOR'2003- 6th Workshop on Computational Processing of the Portuguese Language, pp. 143-150, 2003.

[6] R. Ribeiro and L. C. Oliveira and I. Trancoso, *Morphossyntactic Disambiguation for TTS Systems.*, Proc. of the 3rd Intl. Conf. on Language Resources and Evaluation, v. 5, pp. 1427-1431, 2002.

[7] D. Braga, L. Coelho and F. G. V. Resende Jr., *Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems*, Proceedings of Interspeech 2007, August 27-31, 2007, Antwerp, Belgium.

[8] I. Seara and S. Kafka and S. Klein and R. Seara, *Alternância vocálica das formas verbais e nominais do Português Brasileiro para aplicação em conversão Texto-Fala*, Revista da Sociedade Brasileira de Telecomunicações, n. 1, v. 17, pp. 79-85, 2002.

[9] I. Seara and S. Kafka and S. Klein and R. Seara, *Considerações sobre os Problemas de Alternância Vocálica das Formas Verbais do Português Falado no Brasil para Aplicação em um Sistema de Conversão Texto-Fala*, Anais do XIX Simpósio Brasileiro de Telecomunicações, 2001.

[10] L. Ferrari and F. Barbosa and F. G. V. Resende Jr., *Construções gramaticais e sistemas de conversão texto-fala: o caso dos homógrafos.*, Proc. of the International Conference on Cognitive Linguistics, 2003, Braga, Portugal.

[11] F. Barbosa and L. Ferrari and F. G. V. Resende Jr., *A methodology to analyze homographs for a Brazilian Portuguese TTS system*, PROPOR'2003 - 6th Workshop on Computational Processing of the Portuguese Language, Springer-Verlag, 2003.

[12] F. Barbosa and L. Ferrari and F. G. V. Resende Jr., *A distinção entre homógrafos heterófonos em sistemas de conversão texto-fala*, Processamento da Linguagem, Cultura e Cognição: estudos de linguística cognitiva, 2003, Braga, Portugal.

[13] R. S. Maia and H. Zen and K. Tokuda and T. Kitamura and F. G. V. Resende Jr., *A HMM-based Brazilian Portuguese Speech Synthesizer and its Characteristics.*, Revista da Sociedade Brasileira de Telecomunicações, n. 2, v. 21, pp. 58-71, aug 2006.

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

[15] *Corpus de Extractos de Textos Eletrônicos NILCS/Folha de São Paulo*, Corpus de Extractos de Textos Eletrônicos NILCS/Folha de São Paulo (CETEN-Folha), http://acdc.linguateca.pt/cetenfolha, acessado em 23/02/2008.

[16] A Bilia Sagrada, http://www.culturabrasil.pro.br/zip/biblia.rtf, acessado em 16/03/2009.

[17] Dom Casmurro - Machado de Assis, http://www.machadodeassis.org.br, acessado em 16/03/2009.

[18] The JSpell Project, http://natura.di.uminho.pt/wiki/index.cgi?jspell, acessado em 23/02/2008.

[19] Wordnet, http://wordnet.princeton.edu, acessado em 23/02/2008.