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

Desambiguador de Homógrafos Heterófonos para Sistemas de Conversão Texto-Fala em Português

Daniela Braga e Maria Aldina Marques

Microsoft Language Development Center (MLDC), Universidade do Minho

i-dbraga@microsoft.com, mamarques@ilch.uminho.pt

Abstract: In this paper, a tool for homograph disambiguation in Portuguese Text-to-Speech (TTS) is proposed. This tool works with a part-of-speech parser, used to disambiguate homographs that belong to different parts-of-speech, and a semantic analyzer, used to disambiguate homographs which belong to the same part-of-speech.

This linguistically rule-based methodology will be soon adapted to Brazilian Portuguese, since it involves very little changes. In future work, it is our goal to apply it to other Iberian Languages, such as Galician, Catalan and Mirandese.

The proposed algorithms are meant to solve a significant part of homograph ambiguity in European Portuguese (82 pairs of homographs so far). This system is ready to be integrated in a Letter-to-Sound converter. The algorithms were trained with three different corpora (CETEMPúblico, COMPARA and EUROPARL-Opus) and tested with Natura-Diário do Minho corpus. The obtained experimental results gave rise to 96.9% of accuracy rate.

This paper is structured as following: in section 1, an introduction and state-of-the-art is done; in section 2, the architecture of the homograph disambiguation system in articulation with a TTS system is described; in section 3, the methodology used in the construction of homograph disambiguation algorithms is explained; in section 4, test results are shown and discussed; in section 5, main conclusions and future work are presented.

Palavras-Chave: Síntese de Voz, Conversão Texto-Fala, Text-to-Speech, homógrafos heterófonos, desambiguação, análise morfossintáctica, análise semântica.

1. Caracterização do Problema e Estado da Arte

Tal está, morta, a pálida donzela,
Secas do rosto as rosas e perdida
A branca e viva cor, co a doce vida.
(Camões, *Os Lusíadas*, III, 134)

Este breve excerto saído do trágico episódio do assassínio de Inês de Castro, lapidarmente imortalizado por Camões, contém duas palavras, <secas> e <cor>, cuja decisão de pronúncia depende do conhecimento morfossintáctico e semântico respectivamente. Sem essa informação, nem o falante nem um sintetizador de fala poderá decidir se deve ler a palavra com vogal tónica aberta ou fechada.

A ambiguidade dos homógrafos heterófonos, exemplificada em pares do tipo <o acerto> [e][1] e <eu acerto> [E]; <o almoço> [o] e <eu almoço> [O], ou <eu/ele fora> [o] e <lá fora> [O], representa um problema de difícil resolução nos sistemas de conversão Texto-Fala, sendo responsável por uma considerável taxa de erro.
O que acontece é que a transcrição ortográfica automática, independentemente da abordagem que esteja a ser utilizada, produz erros, já que gera apenas um output (um fone) para cada input (grafema ou conjunto de grafemas), embora, no caso dos homógrafos, devesse ser capaz de escolher entre dois outputs, consoante o contexto morfossintáctico ou semântico do homógrafo em causa.
O problema da desambiguação de homógrafos é realmente complexo porque depende de informação morfossintáctica na maior parte dos casos. Nos pares <o gosto>[o]/ <eu gosto>[O], a diferença de timbre da vogal tónica correlaciona-se com o facto de as palavras pertencerem à classe gramatical de nome e verbo, respectivamente.
Por vezes, a desambiguação de homógrafos só pode ser feita com recurso a informação semântica (entre palavras da mesma categoria gramatical como <sede> [e]/<sede>[E] ou <besta>[e]/ <besta>[E]), sendo esta considerada de mais alto nível e de mais difícil implementação computacional. No excerto seguinte (in Huang et al., 2001:724), saído de uma das obras mais reputadas da actualidade na área do processamento da fala, os autores mostram precisamente que a classificação morfossintáctica da palavra nem sempre é suficiente para determinar a leitura do homógrafo, mesmo para os próprios falantes:

> "Homograph variation can often be solved on POS[2] (grammatical) category. Examples include *object*, *minute*, *bow*, *bass*, *absent*, etc. Unfortunately, correct determination of POS (whether by parsing system or statistical methods) is not always sufficient to resolve pronunciation alternatives. For example, simply knowing that the form *bow* is a noun does not allow us to distinguish the pronunciation appropriate for the instrument of archery from that for the front part of a boat. Even more subtle is the pronunciation of *read* in "If you read the book he'll be angry". Without contextual clues, even human readers cannot resolve the pronunciation of read from the given sentence alone." (Huang et al., 2001: 724)

Ora, enquanto este tipo de conhecimento linguístico vai sendo adquirido e interiorizado pelo ser humano de forma mais ou menos desorganizada desde a infância, através de um processo psico-cognitivo muito complexo, o mesmo não acontece com o computador, que necessita de uma metodologia de aprendizagem muito controlada e estruturada.
A dificuldade inerente a este problema parece explicar a escassez de trabalhos publicados sobre o assunto.

O trabalho de referência sobre a questão da desambiguação de homógrafos em sistemas de TTS é da autoria de David Yarowsky (1996), a partir do qual o autor estabelece uma tipologia de pares de homógrafos para o Inglês, enuncia as várias técnicas tradicionalmente utilizadas para resolver a questão da desambiguação de homógrafos

[1] A transcrição fonética usada neste trabalho segue o alfabeto SAMPA (Computer Readable Phonetic Alphabet) para Português (disponível em: http://www.phon.ucl.ac.uk/home/sampa/portug.htm), acrescido da extensão [l*] para representar a lateral velarizada em situação implosiva presente na palavra <mal>.
[2] POS (Part-of-Speech), também usado por Lyons (1977) equivale a "categoria gramatical" ou "categoria morfológica" da palavra ou ainda"partes do discurso", na Gramática Tradicional.

(N-Gram taggers, classificadores Bayesianos e árvores de decisão) e propõe um algoritmo híbrido, que combina o melhor dos três paradigmas previamente descritos.
Dos principais artigos publicados sobre a problemática da desambiguação de homógrafos no Português aplicada a sistemas TTS, destacam-se as propostas de Ribeiro et al. (2002, 2003) para o PE, Seara et al. (2001, 2002) e Barbosa et al. (2003) e Ferrari et al. (2003) para o PB.
Os trabalhos de Ribeiro et al. (2002, 2003) não se debruçam especificamente sobre o problema da desambiguação de homógrafos, mas antes sobre a influência da informação morfossintáctica no melhor desempenho dos sistemas de TTS e, particularmente na desambiguação de homógrafos heterófonos. No trabalho de 2002, Ribeiro et al. comparam dois analisadores morfológicos, um que segue uma abordagem probabilística e outro que segue uma abordagem híbrida (probabilística e por regras linguísticas). Os resultados parecem mostrar um melhor desempenho global da abordagem híbrida. Apresenta-se ainda uma tabela com uma tipologia de ambiguidades morfossintácticas que influenciam o módulo de análise fonética, ou seja, o conversor grafema-fonema. No entanto, nenhum caso classificado de ambiguidade é acompanhado de exemplos, pelo que não se percebe quando se trata de ambiguidade morfossintáctica decorrente da homonímia, ou ambiguidade morfossintáctica decorrente da homografia heterófona. A actualização do mesmo trabalho publicada em 2003 vem precisamente corroborar que a desambiguação morfossintáctica analisada é, essencialmente, a desambiguação de palavras homónimas, o que tem pouco impacto ao nível dos módulos de conversão grafema-fone dos sistemas de TTS, visto não ter consequências na articulação da palavra. Este trabalho mostra, no entanto, o impacto que a resolução de ambiguidade morfossintáctica pode ter ao nível do módulo de geração prosódica, ao ser capaz de distinguir palavras conteúdo e palavras função, com impacto no foco da frase, e ao possibilitar a delimitação dos grupos prosódicos.
Ferrari et al. (2003) propõem uma metodologia linguística, assente na Gramática Cognitiva, para solucionar a questão da variação fonética dos homógrafos heterófonos, com base na análise de corpora. A análise centra-se na identificação e programação das construções sintácticas vizinhas esperadas, partindo apenas da análise do contexto:

> "Since the nouns [sedi] and [sEdi] can take part in noun phrases, prepositional phrases or verb phrases, the analysis focused on different types of constructional schemas that are relevant for the distinction between them." (Ferrari et al., 2003)

Esta abordagem permite efectuar desambiguação não só morfossintáctica como semântica. Contudo, revela-se pouco económica, dado necessitar de um estudo de ocorrências contextuais análogo para cada par de homógrafos heterófonos, o que não contribui para a desejável programação optimizada dos algoritmos que devem compor o módulo de conversão grafema-fone.
Nos trabalhos de Seara et al. (2001, 2002), desenvolve-se, através de uma abordagem linguística, um *parser* ou analisador morfossintáctico com vista a resolver a questão da alternância vocálica existente em formas nominais e verbais. Trata-se de um trabalho muito interessante e importante para a resolução da ambiguidade presente em alguns tipos de homógrafos heterófonos, por um lado, e de resolução da alternância vocálica ao longo da flexão verbal, como em <eu meto>[e]/<ele mete>[E]. No entanto, este trabalho não abrange os casos em que a desambiguação de homógrafos heterófonos se estabelece semanticamente. No presente trabalho, fizemos uma re-estruturação da tipologia enunciada em Seara et al. (2001, 2002), adaptando-a apenas a casos de homografia

heterófona e aumentando a cobertura dos pares de homógrafos, através da integração da análise semântica.

2. Arquitectura do Sistema

O desambiguador de homógrafos heterófonos constitui uma componente do módulo de Análise e Transcrição Fonética (vide Figura 1), articulando-se directamente com o Conversor Grafema-Fone[3]. Esta componente insere-se na parte que se designa por *front-end* ou pré-processamento do texto e faz a conversão do texto em etiquetas fonéticas, as quais são seguidamente interpretadas pelo motor de síntese ou *back-end*. Uma base de dados de voz ou *voice font*, foneticamente etiquetada, fornece os sons da língua, fones, difones ou outras unidades, que o motor de síntese transforma e faz corresponder às etiquetas fonéticas, gerando assim voz sintética.

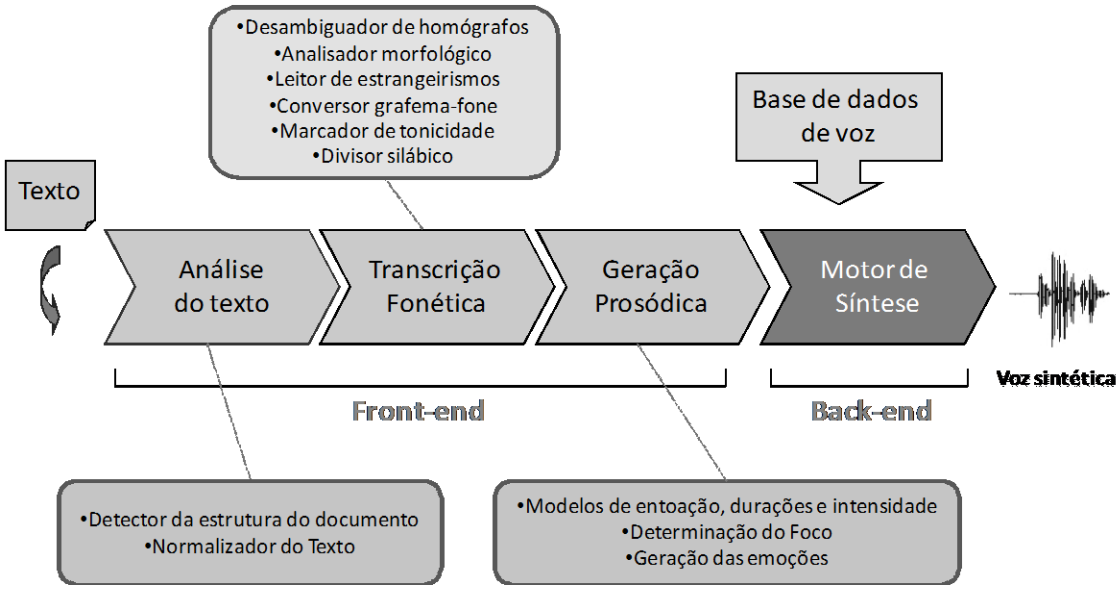

Figura 1: Esquema da arquitectura de um sistema de conversão Texto-Fala.

Porém, cada uma das componentes que fazem parte dos módulos do sistema constitui uma ferramenta complexa, que funciona quer individualmente, quer articuladamente com os outros módulos do sistema.
Na Figura 2, que passaremos a descrever, pode ver-se a estrutura do Desambiguador de Homógrafos Heterófonos. O Desambiguador pode ser encarado como uma parte do Analisador Morfológico. Na verdade, o seu funcionamento está dependente da análise morfossintáctica, como ilustra a Figura 2, por ser necessário indentificar as categorias gramaticais das palavras que ocorrem à esquerda e à direita do homógrafo em análise.

[3] Também designado por Conversor LTS (Letter-to-Sound) ou G2P (Grapheme-to-phone/me).

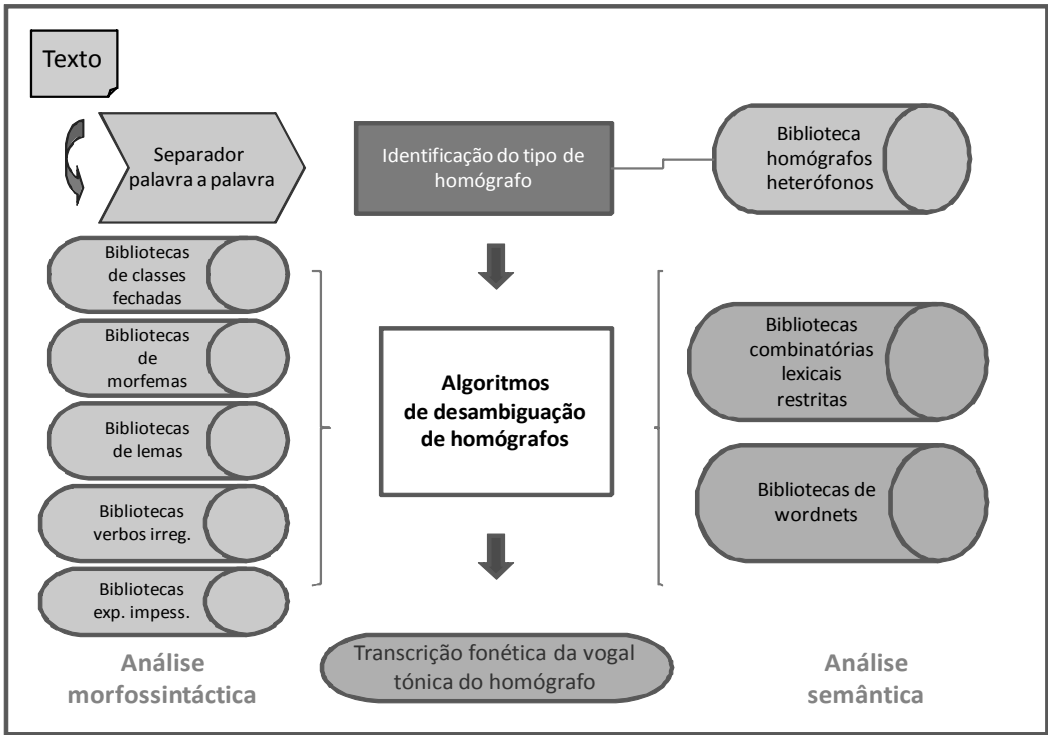

Figura 2: Arquitectura do Desambiguador de Homógrafos Heterófonos.

A partir da observação da Figura 2, pode ver-se que o input do sistema é o texto que é separado palavra a palavra. A seguir, um algoritmo de busca encarrega-se de verificar se existem homógrafos no texto de entrada e de identificar o seu tipo, através da consulta à biblioteca de homógrafos heterófonos. Estão até ao momento identificados 21 tipos de homógrafos, que são apresentados no ponto 3, o que faz com que haja 21 outputs possíveis nesta fase do sistema.
O passo seguinte consiste em fazer passar o homógrafo em questão pelo algoritmo de desambiguação que lhe foi atribuído. Estes algoritmos consistem em árvores de decisão que formulam várias perguntas relativas ao contexto morfossintáctico do homógrafo e que, com base nas respostas, permitem decidir a sua pronúncia, como será descrito no ponto 3. Para determinar a categoria gramatical das palavras vizinhas, o sistema consulta o Analisador Morfossintáctico, que é constituído por várias bibliotecas e por regras morfossintácticas que permitem gerar a classificação gramatical.

Fazem parte do Analisador Morfológico as seguintes bibliotecas:

1. **Biblioteca das classes fechadas**, ou seja, as categorias gramaticais cujos itens lexicais existem em número finito e que dificilmente admitem formação de novas palavras. Nestas bibliotecas não se incluiram as palavras que apresentam homonímia gramatical, como por exemplo <o>, <a>, <os>, <as>, <muito>, <pouco>, <tanto>, <que>, <quem>, <onde>, entre outras. Desta biblioteca foram consideradas as seguintes classes:

    a. preposições (PREP)
    b. advérbios (ADV) e advérbios de quantidade (ADV_Q)
    c. contracções de preposição com determinante/pronome (CONT)

d. conjunções subordinativas (C_S) e locuções conjuncionais subordinativas (Loc_S)
e. conjunções coordenativas (C_C) e locuções conjuncionais coordenativas (Loc_C)
f. determinantes artigos indefinidos (ART_IND)
g. pronomes e determinantes demonstrativos (DEM)
h. pronomes e determinantes possessivos (POSS)
i. pronomes e determinantes indefinidos (IND)
j. pronomes e determinantes interrogativos (INT)
k. numerais (NUM)
l. pronomes pessoais sujeito (P_PES_SU) e pronomes pessoais objecto (P_PES_O_1)[4], (P_PES_O_2)[5], (P_PES_O_3)[6]
m. pronomes relativos (P_REL)
n. interjeições (INTJ)

2. **Biblioteca de afixos**[7], constituída pelas seguintes sub-classes:
   a. Sufixos Nominais (Des_N), Adjectivais (Des_Adj) e Adverbiais (Des_Adv)
   b. Sufixos verbais (Des_V)
   c. Prefixos portugueses (Pref_PT)
   d. Radicais gregos e latinos (R_GL)

3. **Bibliotecas de verbos irregulares,** contendo as formas dos principais verbos irregulares.

4. **Biblioteca de expressões impessoais** (Exp_Imp), contendo expressões constituídas por verbo ser na 3ª pessoa seguido de adjectivo (ex: <é importante>). Estas expressões regem orações completivas integrantes ou infinitivas, permitindo assim prever a sua sintaxe.

5. **Biblioteca de lemas**[8]**,** constituída pelo dicionário Jspell[9] para o Português, com cerca de 34000 palavras, anotado morfologicamente, que resultou do projecto

[4] Pronomes pessoais objecto que não sofram processos de assimilação resultantes da co-articulação com formas verbais (ex: <me>, <te>, <se>, <lhe>...).
[5] Pronomes pessoais objecto na terceira pessoa que sofrem assimilação no contacto com formas verbais com <–r>, <-s>, e <–z> em situação implosiva (ex: <vou comprá-lo>).
[6] Pronomes pessoais objecto na terceira pessoa que sofrem assimilação no contacto com formas verbais com nasal ou ditongo em situação implosiva (ex: compram-no).
[7] Entendemos o conceito de afixo como constituinte morfológico que se associa ao radical e tema, os constituintes básicos da palavra, segundo uma perspectiva inovadora da Teoria X-Barra aplicada à Morfologia do Português por Alina Villalva: "No Português, os afixos disponíveis são prefixos, quando ocorrem na periferia esquerda da forma de base (...), e sufixos, quando se encontram à direita da forma de base." (in Mateus, et al. 2003: 941).
[8] Segundo Iriarte (2001: 30) "O lema (entrada ou vedeta) poderá ser qualquer palavra, conjunto de palavras, signo, letra, conjunto de letras ou signos que encabeçam um artigo de dicionário, enciclopédia, índice, ficha, etc., e que é objecto de definição, explicação, tratamento enciclopédico ou, no caso dos dicionários bilingues, do qual se fornece um equivalente noutra língua (...)". Neste sentido, o lema pode corresponder a uma palavra (ex: hierro), uma sigla (ex: DNI) ou um sintagma (ex: caminho de ferro) (Iriarte, 2001:300). Na nossa biblioteca corresponde apenas a palavras.
[9] Sobre o Jspell: "O Jspell é um analisador morfológico *open source* para sistemas baseados em UNIX, baseado no Ispell, que permite mediante diversos tipos de interface analisar morfologicamente ou corrigir

Natura, ainda em curso[10], levado a cabo por investigadores do pólo de Braga da Linguateca[11], José João Almeida, Alberto Simões e Rui Vilela.

Contam-se entre as principais fontes para a constituição das bibliotecas de classes fechadas e morfemas as obras de Cunha & Cintra (1992), Estrela et al. (2004) e Bergström et al. (1997). As bibliotecas de verbos irregulares partiram da lista de verbos irregulares disponível no pacote Jspell, estando em processo de ampliação manual com apoio bibliográfico (Nogueira, 1994).

A identificação do homógrafo faz-se através da consulta à **Biblioteca de homógrafos**, que ainda está em fase de expansão. Esta biblioteca contém 82 lemas com a informação do tipo de homógrafo a que pertencem, a que corresponde um algoritmo de decisão. Se a palavra em questão estiver na lista de homógrafos, é encaminhada para o seu respectivo algoritmo de decisão.

A análise morfossintáctica ocorre sempre que os homógrafos pertençam a classes gramaticais distintas. Neste caso, consultam-se as bibliotecas da Figura 2 que são responsáveis pela análise morfossintáctica do texto.

Mas sempre que os pares de homógrafos pertençam à mesma categoria gramatical, a análise morfológica dá lugar à análise semântica, operada pela consulta das bibliotecas de combinatórias lexicais restritas[12] e as bibliotecas de *Wordnets*, cujo funcionamento será descrito no ponto 3.

As bibliotecas de combinatórias lexicais restritas abrangem, segundo a designação de Iriarte (2001), os frasemas completos (ou expressões idiomáticas)[13] (ex: "cor de burro quando foge".), os semi-frasemas (ou colocações)[14] (ex: "pregar um susto") e os quase-

---

a ortografia de um texto. Está orientado para o processamento de textos/palavras da língua portuguesa. (...) O Jspell está disponível para língua portuguesa, inglesa, e latim, sob a licença GNU GPL2.Os dicionários não pretendem cobrir "todo" o vocabulário existente, apenas as formas mais frequentes. As palavras cuja terminologia é demasiado específica e raras, não são incluídas no dicionário. (...) O dicionário para o Português (1995), morfologicamente anotado, foi construído a partir da extracção de palavras de material académico da Universidade do Minho, como teses de Doutoramento e Mestrado, corpora jornalístico Português de Portugal disponível publicamente, listas de nomes públicas, e diverso material livre de direitos de autor. Numa segunda fase, modificações individuais consoante o critério dos autores, recurso à consulta de prontuários, dicionários de definições, lista de frequências, sugestões de utilizadores, cruzamento e validação de palavras com a colaboração de entidades externas." (in: http://linguateca.di.uminho.pt/jspell/jsolhelp.pl)

[10] Disponível para download em: http://natura.di.uminho.pt/wiki/index.cgi?jspell.

[11] Para mais informação sobre o pólo de Braga da Linguateca, consultar http://linguateca.di.uminho.pt/.

[12] A combinatória lexical restrita é uma unidade pluriverbal em que os seus elementos se combinam para produzir um determinado sentido e que, se forem truncadas ou um dos seus elementos substituídos, se torna agramatical. Iriarte, a propósito dos sintagmas"mudança radical", "dar um passeio" e "leite gordo", explica: "É evidente que este tipo de combinações lexicais não são totalmente livres, como fica evidenciado pelos casos agramaticais que acompanham cada exemplo (*fazer um passeio, etc.). Estamos perante casos de combinatória lexical restrita (as chamadas colocações) (...)"(Iriarte: 2001: 139).

[13] Segundo Iriarte (2001: 174): "Uma expressão idiomática ou frasema completo AB ("ser o braço direito de") é uma combinação de dois ou mais lexemas A (braço)e B (direito), cujo significante é a sima regular dos significantes dos lexemas constituintes /A+B/ (braço+direito), mas cujo significado não é a esperada união regular de A e B (...), mas um significado diferente 'C' ('[ser o] auxiliar principal' ou 'principal colaborador') que não inclui nem 'A' nem 'B'.".

[14] Ainda segundo o mesmo autor: "(...) uma colocação, ou semi-frasema, AB é uma combinação de dois ou mais lexemas A e B, cujo significante é a soma regular dos dois significantes dos lexemas constituintes /A + B/, e cujo significado 'X' inclui o significado do lexema A mais um significado 'C' ('X'='A+C') , de tal maneira que o lexema B que exprime 'C' não é seleccionado livremente. Numa colocação, pensemos, por exemplo, em ódio mortal, um dos seus constituintes, A (ódio), é seleccionado pelo falante por causa do seu significado, que é conservado intacto; mas o segundo elemento constituinte

frasemas[15] (ex: "boca do lobo"). Destas bibliotecas constam ainda os provérbios (ex: "Gato escaldado de água fria tem medo"). Estas bibliotecas foram construídas, para cada par de homógrafo, a partir da análise de vários corpora electrónicos, designadamente o corpus jornalístico do CETEM-Público[16], o COMPARA[17] (corpus paralelo em Português e em Inglês constituído por textos literários e suas traduções) e o EUROPARL – Opus[18] (constituído por transcrições dos debates do parlamento europeu; corpus alinhado para 12 línguas).
As bibliotecas de Wordnets[19], obtidas pelo mesmo processo que as anteriores, procuram reunir palavras semanticamente relacionadas previsíveis de co-ocorrerem com a palavra a que se ligam. A cada homógrafo com a mesma categoria gramatical são-lhe associadas uma biblioteca de combinatórias lexicais restritas e uma Wordnet.

3. Algoritmos de Desambiguação de Homógrafos Heterófonos

3.1. Metodologia

Este trabalho foi iniciado com uma recolha exaustiva de pares de homógrafos em todas as fontes bibliográficas encontradas, desde gramáticas prescritivas a prontuários, visto que o bom desempenho do nosso desambiguador depende da presença do homógrafo em análise na nossa biblioteca de homógrafos. No entanto, neste tipo de bibliografia, os homógrafos são tratados sempre da mesma forma e usando sempre os mesmos exemplos clássicos. A nossa lista foi assim sendo ampliada através de sucessivos testes ao conversor grafema-fone (Braga et al., 2006), ainda em desenvolvimento, que nos permitiram identificar os actuais 82 homógrafos[20] que compõem a nossa lista até à data de redacção do presente trabalho.
A fase seguinte consistiu na organização dos homógrafos por tipos, de acordo com a natureza da sua oposição e com a alternância fonética que continha. A cada tipo fez-se

---

B (mortal), significa 'C' ('intenso'), diferente de 'B' (que causa ou pode causar a morte). Fora da colocação AB, B (mortal) não seria usado para exprimir C ('intenso') (...)" (Iriarte, 2001: 176)

[15] Os quase frasemas "são frasemas em que, para além de se conservarem os sentidos dos lexemas que os constituem, acrescenta-se um novo sentido que não é dedutível da simples soma dos sentidos dos lexemas constituintes(...). São exemplos de quase-frasemas, tecto falso (...), onde para além dos sentidos 'tecto' e 'falso' temos também o sentido 'para isolar acústica e termicamente' (...)" (Iriarte, 2001: 181-182).

[16] Disponível em: http://www.linguateca.pt/.

[17] Disponível em: http://www.linguateca.pt/COMPARA/BuscaSimples.html

[18] Disponível em: http://logos.uio.no/cgi-bin/opus/opuscqp.pl?corpus=EUROPARL;lang=pt

[19] O conceito de Wordnet surgiu da designação de uma base de dados de palavras, construída para o Inglês sob direcção de George A. Miller, constituída por palavras (nomes, verbos, adjectivos, advérbios) agrupadas por relações semânticas de base cognitiva, cada uma expressando um conceito. Cada palavra cria uma rede de outras palavras e conceitos, através da qual é possível navegar. Trata-se de um recurso muito útil para o processamento da linguagem natural. A Wordnet é open source e está disponível em: http://wordnet.princeton.edu/.Está em curso o projecto de criação de uma Wordnet para o Português , no Centro de Linguística de Lisboa (http://www.clul.ul.pt/clg/projectos/WordNet.PT-I.html), mas com resultados ainda não disponibilizados.

[20] Segue-se a lista obtida até à redacção deste trabalho, ordenada alfabeticamente: aborto, aceno, acerto, acordo, adorno, almoço, apelo, aperto, apreço, arrepelo, arrojo, arroto, besta, boto, cerca, choco, choro, colher, começo, concerto, conserto, consolo, contorno, cor, coro, corte, desemprego, despojo, desses, deste, destes, destroço, emprego, encosto, erro, esforço, espeto, este, estorvo, folgo, fora, forma, gelo, golo, gosto, governo, interesse, interesses, jogo, leste, lobo, logro, medo, meta, modelo, molho, namoro, olho, pegada, peso, piloto, pregar, reforço, rego, remo, rogo, rola, rolo, seco, sede, selo, sobre, soco, sopro, suborno, termos, testo, toco, topo, torno, troco, troço.

corresponder um algoritmo de decisão. A nível da implementação, verificou-se que os algoritmos podiam ser agrupados em menos tipos, uma vez que o conjunto de perguntas era o mesmo, por exemplo, para homógrafos que pertencessem à mesma categoria gramatical, mudando apenas a saída fonética, tal como acontece com os algoritmos 1 e 2. São também muito semelhantes os algoritmos cuja saída é verbo, sendo que a única alteração se verifica em pequenos detalhes, consoante o homógrafo é uma forma verbal na 1ª ou na 3ª pessoas do Presente do Indicativo (ex: <gosto> e <rola>).
Seguidamente, procedeu-se à elaboração de regras sintácticas de desambiguação de homógrafos. Este processo foi acompanhado de buscas electrónicas em corpora, no sentido de validar e consolidar as nossas intuições linguísticas.
Usaram-se o CETEM-Público (corpus jornalístico), o COMPARA (corpus literário) e do EUROPARL – Opus (corpus de debate parlamentar). Esta diversidade de corpora pareceu-nos importante para encontrar mais concordâncias em contexto e contextos mais diversificados decorrentes dos diferentes tipos de texto. Cada homógrafo foi inserido no sistema de busca disponibilizado. O sistema apresentou, em seguida, o número e as ocorrências da palavra em contexto, permitindo assim confirmar regras e verificar mais casos.
Finalmente, os algoritmos foram implementados e o seu desempenho foi testado, como se descreverá no ponto 4.

3.2. Tipologia de homógrafos heterófonos

Nas tabelas 1 e 2 que se seguem, apresentam-se as tipologias de homógrafos consideradas.
Na tabela 1, estão listados os homógrafos cuja desambiguação se estabelece pela identificação da categoria gramatical da palavra.
Os tipos 1 e 2 são os que encerram maior número de pares, uma vez que, em Português, a maior parte dos homógrafos ocorre em oposições de Nome masculino singular *versus* Verbo na primeira pessoa gramatical do Presente do Indicativo. Do total de 82 pares de homógrafos, 54 pertencem aos tipos 1 e 2, ou seja, 66% do total de homógrafos. Estes dois primeiros tipos apresentam algoritmos de desambiguação iguais, diferindo apenas na saída fonética.
A oposição gramatical mais produtiva é, assim, a que opõe Nome a Verbo, presente também nos tipos 3, 4, 8, 12, 13 e 14, embora os dois últimos apresentem uma alternância tripartida, uma vez que o homógrafo pode desempenhar três funções gramaticais.
Do ponto de vista da alternância vocálica, as oposições mais produtivas são as que se estabelecem ao nível da vogal do radical, opondo sistematicamente as vogais orais semi-fechadas [e] e [o] às vogais orais semi-abertas [E] e [O], respectivamente.
De salientar, é o facto de nos Nomes as vogais do radical serem frequentemente semi-fechadas, ao passo que nas formas verbais elas se tornam invariavelmente semi-abertas.
O tipo 13 é um caso particularmente complexo de desambiguação, porque necessita de análise semântica para a oposição <f<u>o</u>rma> [o] e <f<u>o</u>rma> [O], que se trata da mesma categoria gramatical, e de análise morfológica para distinguir estas palavras da correspondente forma flexionada do verbo na terceira pessoa do singular do Presente do Indicativo. Como se pode depreender da análise da Tabela 1, outras oposições gramaticais (tipos 5, 6, 9, 11) e vocálicas (tipo 12) são possíveis também.

É ainda de destacar o facto de os homógrafos de tipo 12 não apresentarem alternância na vogal tónica, mas sim na vogal pré-tónica.
A Tabela 2 exibe os pares de homógrafos cuja desambiguação se estabelece por critérios semânticos, recorrendo portanto a bibliotecas de combinatórias lexicais restritas e bibliotecas de Wordnets, que foram constituídas através da análise dos corpora supra mencionados.

Tipo	Categoria gramatical da oposição	Exemplo	Alternância vocálica da oposição
1	Nome (masc. sing.) /Verbo (1ª p.s. Pres. Ind.)	O erro foi corrigido. Eu erro muito.	[e] Nome / [E] Verbo
2	Nome (masc. sing.) /Verbo (1ª p.s. Pres. Ind.)	Tens um olho vermelho. Eu olho para ele muitas vezes.	[o] Nome/ [O] Verbo
3	Nome (fem. sing.) /Verbo (3ª p.s. Pres. Ind.)	Vi uma rola branca. A pedra rola.	[o] Nome/ [O] Verbo
4	Verbo/Nome	Vou colher morangos. Falta-me uma colher.	[e] Verbo / [E] Nome
5	Contracção/Verbo	Quero um desses. Se tu me desses um beijo...	[e] Contracção / [E] Verbo
6	Verbo/ Advérbio	Antes ele fora médico. Lá fora está frio.	[o] Verbo/ [O] Advérbio
7	Adj. ou Nome/ Verbo(1ª p.s. Pres. Ind.)	O pão está seco. Eu seco a roupa no estendal.	[e] Adj., N/ [E] Verbo
8	Adj. ou Nome/ Verbo(1ª p.s. Pres. Ind.)	Tem o pé boto[21]? Eu boto sal nos bolos.	[o] Adj., N/ [O] Verbo
9	Demonstrativo/ Nome ou Adj.	Este carro é meu. Norte, sul, este, oeste.	[e] Dem. / [E] Adj., N
10	Verbo /Adj. ou Nome	Leste o anúncio? Fica virado a leste.	[e] Verbo / [E] Adj., N
11	Preposição/Verbo	Falou sobre a vida. Espero que não sobre.	[o] Prep./ [O] Verbo
12	Verbo ou Adj./ Nome	Foi conversa pegada. Vi pegadas de dinossauros.	[@] Verbo./ [E] Nome
13	Nome/Nome/Verbo	Tira o bolo da forma. Essa forma é circular. Ele forma os alunos.	[o] Nome/ [O] Nome/ [O] Verbo
14	Prep./ Nome/ Verbo	Vi cerca de dois lobos. A cerca é de madeira. Ele cerca o castelo.	[e] Prep./ [e] Nome/ [E]Verbo

Tabela 1: Tipos de homógrafos pertencentes a classes morfossintácticas diferentes.

A oposição gramatical é essencialmente estabelecida entre Nomes com significados diferentes (excepto no tipo 21, que opõe Verbo a Verbo), ao passo que a alternância vocálica ocorre sistematicamente entre as vogais orais semi-fechadas [e] e [o] e as vogais orais semi-abertas [E] e [O], respectivamente. Seja como for, a estratégia de análise semântica foi também utilizada para auxiliar na desambiguação de homógrafos de tipo 6, 13 e 14, uma vez que a análise morfossintáctica se revelou insuficiente.
A desambiguação semântica é feita caso a caso, uma vez que a cada par de homógrafos corresponde um algoritmo de decisão separado.

[21] <boto> tem mais significados, embora pouco usuais. <boto> /O/ é s.m. em Portugal, como regionalismo de Trás-os-Montes, significando "borracha"; é s.m. /o/, dando nome a um tipo de cetáceos da família dos delfinídeos". Em Bras./Gir. significa "coisa volumosa". Também como s.m. /o/ é um sacerdote do hinduísmo (in Dicionário da Academia das Ciências de Lisboa).

Tipo	Categoria gramatical da oposição	Exemplo	Alternância vocálica da oposição
15	Nome/Nome	A b<u>e</u>sta foi abatida. Manejava bem a b<u>e</u>sta.	[e] Nome/ [E] Nome
16	Nome/Nome	Tinha muita s<u>e</u>de. A s<u>e</u>de da ONU...	[e] Nome/ [E] Nome
17	Nome/Nome	Tem m<u>e</u>do de voar. Os m<u>e</u>dos são um povo.	[e] Nome/ [E] Nome
18	Nome/Nome/Verbo	Há café na t<u>e</u>rmos. É bom t<u>e</u>rmos saúde.	[e] Nome/ [E] Nome/ [E]Verbo
19	Nome/Nome	O tecido perdeu a c<u>o</u>r. Sei isso de c<u>o</u>r e salteado.	[o] Nome/ [O] Nome
20	Nome/Nome	O l<u>o</u>bo é um animal. Tem lesões no l<u>o</u>bo occipital.	[o] Nome/ [O] Nome
21	Verbo/Verbo	Vou pr<u>e</u>gar um prego. Vai pr<u>e</u>gar aos peixinhos.	[@] Verbo/ [E] Verbo

Tabela 2: Pares de homógrafos com a mesma classe morfossintáctica.

3.3. Algoritmos de desambiguação

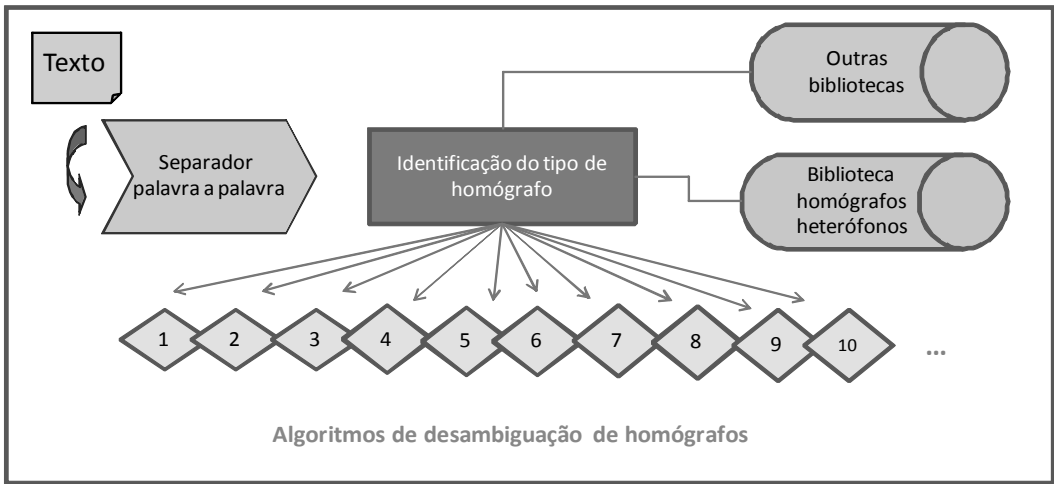

Figura 3: Funcionamento do Desambiguador de Homógrafos Heterófonos.

Após a separação do texto em palavras estar efectuada, o desambiguador começa por buscar candidatos a homógrafos por consulta à sua biblioteca de homógrafos. Se o sistema identificar uma dada palavra como homógrafo, procede à identificação do tipo a que pertence, para em seguida lhe associar um dado algoritmo que permitirá prever o output fonético (*vide* Figura 3).

Na Tabela 3, apresentam-se os símbolos usados na representação gráfica dos algoritmos, bem como o seu significado.

Por limitações de espaço, apresentamos apenas alguns algoritmos para ilustrar o funcionamento da nossa metodologia.

Após a identificação do homógrafo com o seu tipo, o sistema submete-o a várias perguntas relativas às palavras que com ele co-ocorrem à esquerda e à direita.

Seguimos duas estratégias no desenho dos nossos algoritmos. Nuns casos, no primeiro losango, surge uma bateria de perguntas com o objectivo de conduzir à saída mais provável após análise dos corpora. Se a resposta for negativa, então passa-se para o segundo losango, contendo as perguntas que conduzirão à saída estatisticamente menos

provável. São exemplos deste funcionamento, os algoritmos 1, 3, 14, 16 ou 21 (*vide* Figuras 4, 5, 8, 9, 10).

Símbolo	Significado
P-1, P-2, P-3	última, penúltima e antepenúltima palavras, respectivamente
P+1, P+2, P+3	primeira, segunda e terceira palavras seguintes, respectivamente
F-1, F-2, F-3	última, penúltima e antepenúltima frases, respectivamente
F0	a própria frase
F+1, F+2, F+3	primeira, segunda e terceira frases seguintes, respectivamente
DEM	pronome ou determinante demonstrativo
IND	pronome ou determinante indefinido
INT	pronome ou determinante interrogativo
POSS	pronome ou determinante possessivo
ART_IND	artigo indefinido
P_REL	pronome relativo
PREP	preposição
CONT	contracção da preposição com determinante
P_PES_S, P_PES_O	pronome pessoal sujeito, pronome pessoal objecto
CONJ_S, CONJ_C	conjunção subordinada, conjunção coordenada
Loc_S, Loc_C	locuções conjuncionais subordinativa e coordenativa
ADV, ADV_Q	advérbio, advérbio de quantidade
NUM	numeral
DIG	dígito
INTJ	interjeição
Des_V	desinência ou sufixo verbal
PART	particípio
Des_N	desinência ou sufixo nominal
Des_Adj	desinência ou sufixo adjectival
Des_Adv	desinência ou sufixo adverbial
Pref_PT	prefixo português
R_GL	radical grego ou latino
ends by	que termine por
P_M	palavra ou expressão começada por maiúscula
+	seguido de
,	ou
or	condição alternativa
and	condição aditiva

Tabela 3: Simbologia usada nos algoritmos.

Em outros casos, há apenas uma bateria de perguntas com duas saídas, caso a resposta seja afirmativa ou negativa. A resposta positiva corresponde à saída menos provável. Se a resposta for negativa, obtemos a saída mais provável. São exemplos deste funcionamento, os algoritmos de tipo 5 e 7 (*vide* Figuras 6 e 7). Muitos outros contextos estarão em falta, certamente. No entanto, o desenho dos algoritmos baseou-se nos tipos de ocorrências encontradas nos corpora disponíveis, assegurando pelo menos os contextos estatisticamente mais representativos.

Apesar da eficácia dos algoritmos, ainda há casos de ambiguidade lexical que nem por análise semântica são facilmente resolúveis, como neste excerto em que ocorre um homógrafo de tipo 2:

> "Depois, se tal palavra tem algum sentido aplicada a um quebrantamento que não durou mais que uns instantes, e já naquele estado de meia vigília que vai preparando o despertar, considerou seriamente que não estava bem manter-se numa tal indecisão, **acordo, não acordo, acordo, não acordo**, sempre chega uma altura em que não há outro remédio que arriscar."
>
> (COMPARA, PPJSA1 (116)).

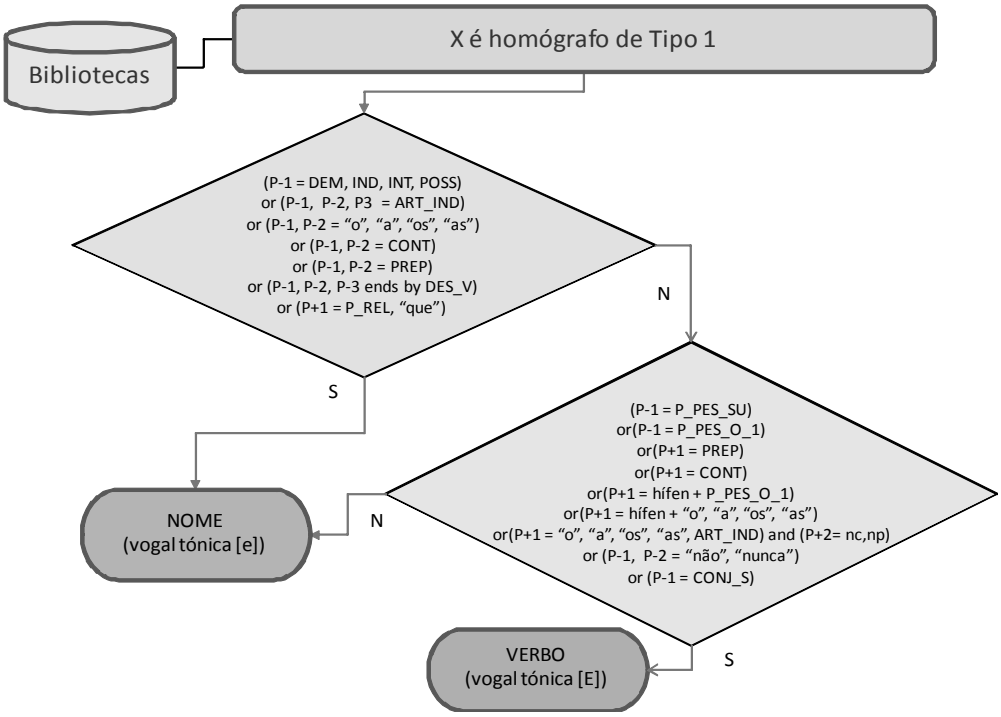

Figura 4: Algoritmo de desambiguação de homógrafos de tipo 1 (ex: 'apelo').

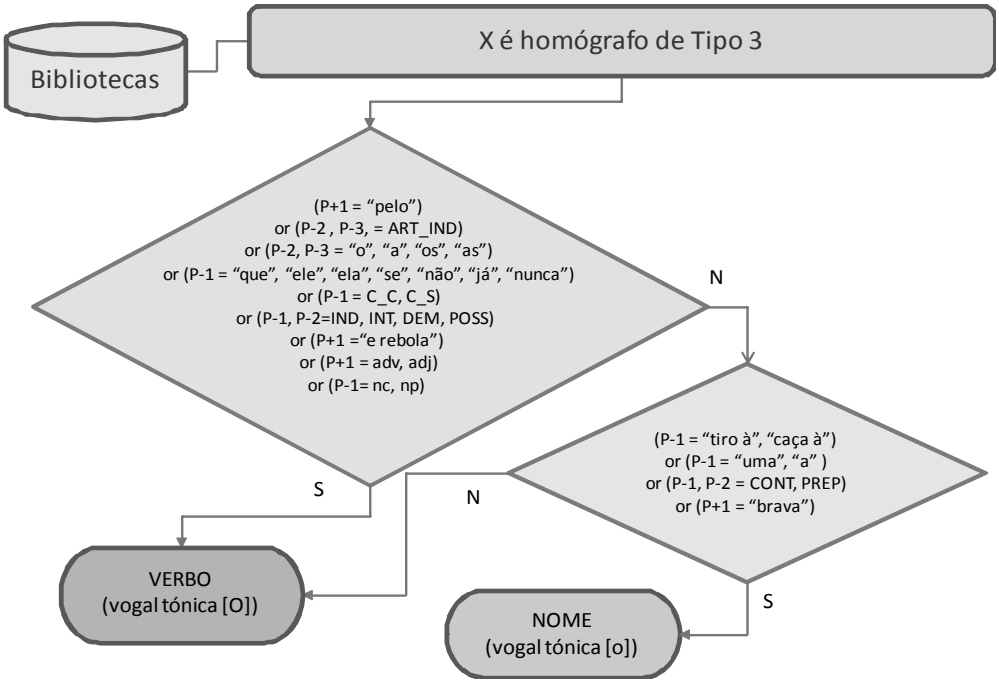

Figura 5: Algoritmo de desambiguação de homógrafos de tipo 3 ('rola').

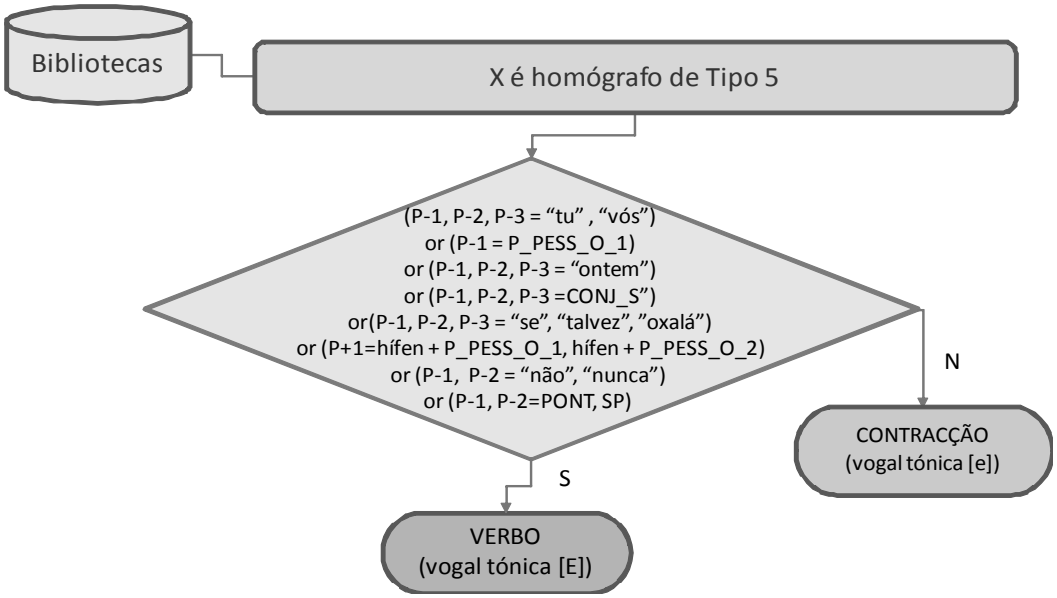

Figura 6: Algoritmo de desambiguação de homógrafos de tipo 5 (ex.’deste’).

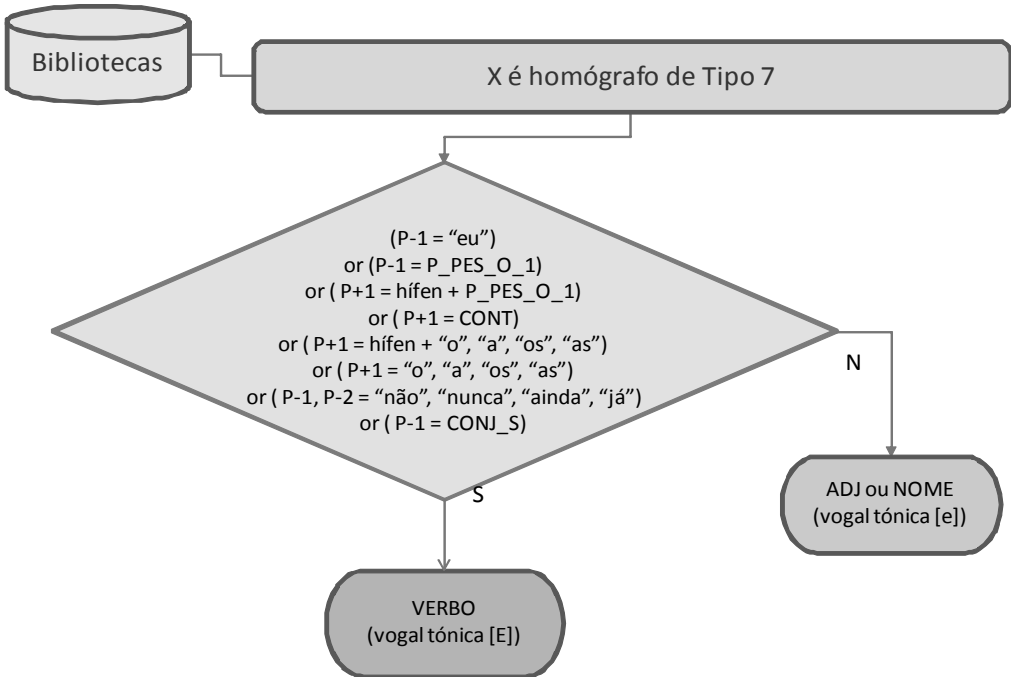

Figura 7: Algoritmo de desambiguação de homógrafos de tipo 7 (’seco’).

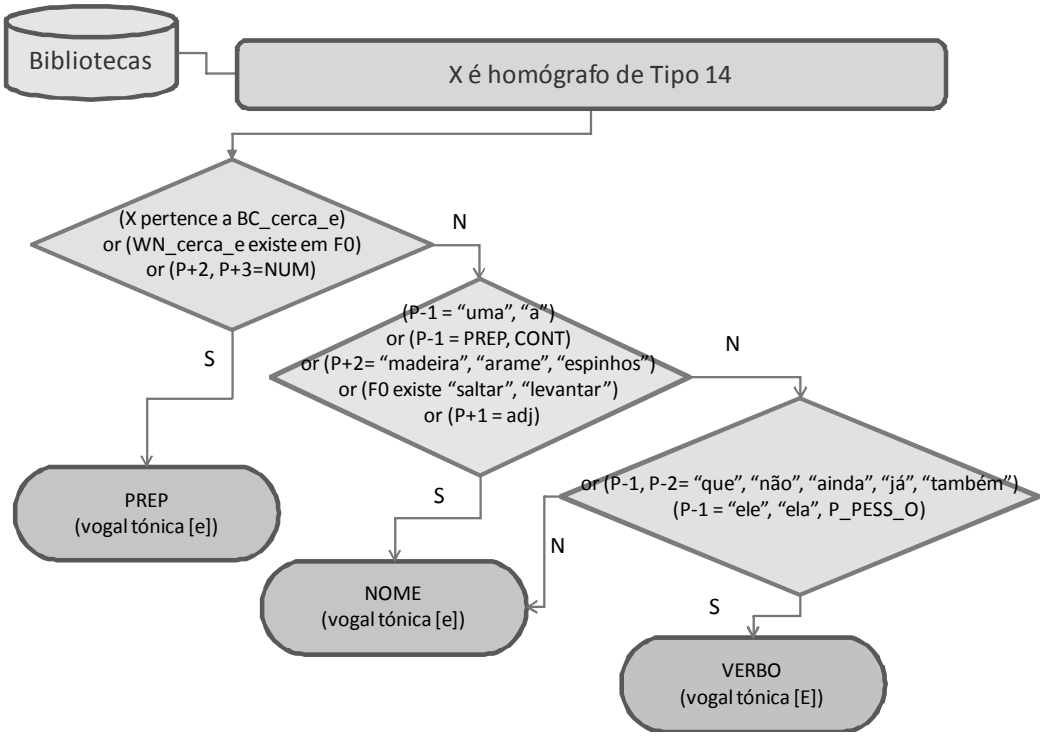

Figura 8: Algoritmo de desambiguação de homógrafos de tipo 14 ('cerca').

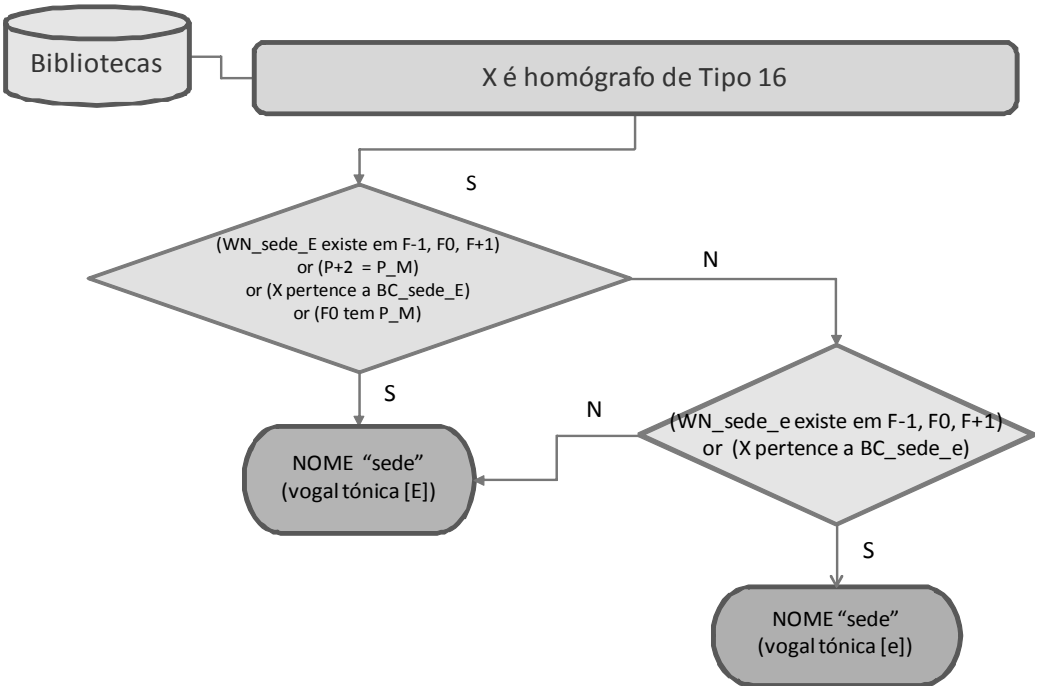

Figura 9: Algoritmo de desambiguação de homógrafos de tipo 16 ('sede').

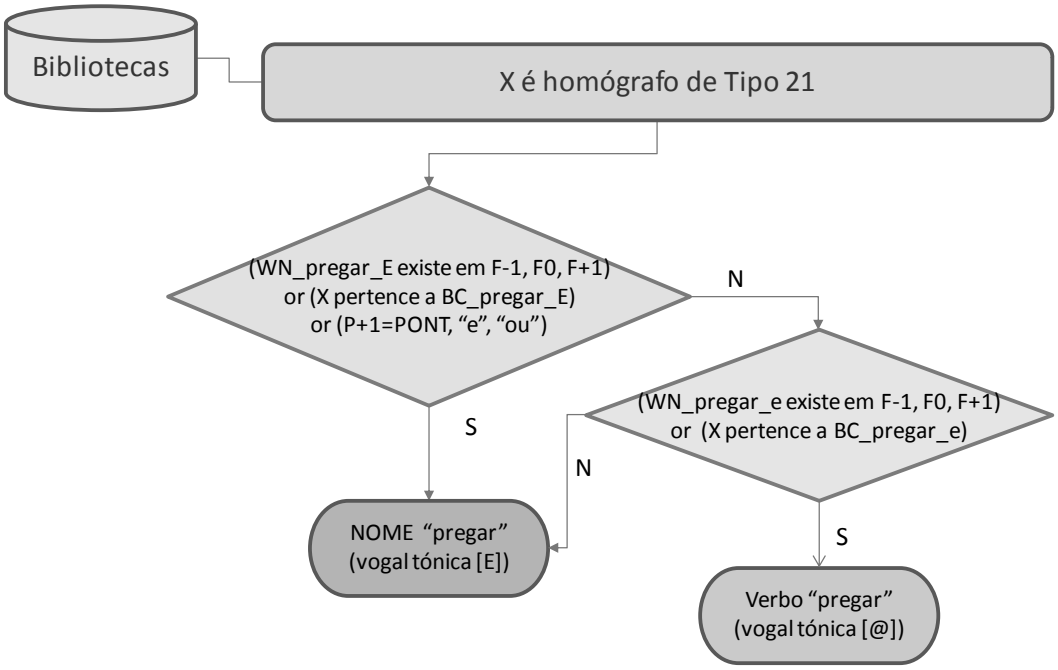

Figura 10: Algoritmo de desambiguação de homógrafos de tipo 21 ('pregar').

4. Testes e discussão de resultados

Ao nível da implementação, foi possível reduzir o número de algoritmos, uma vez que a uma dada categoria gramatical corresponde um certo conjunto de perguntas.
Na Figura 11, pode ver-se a *interface* do sistema.

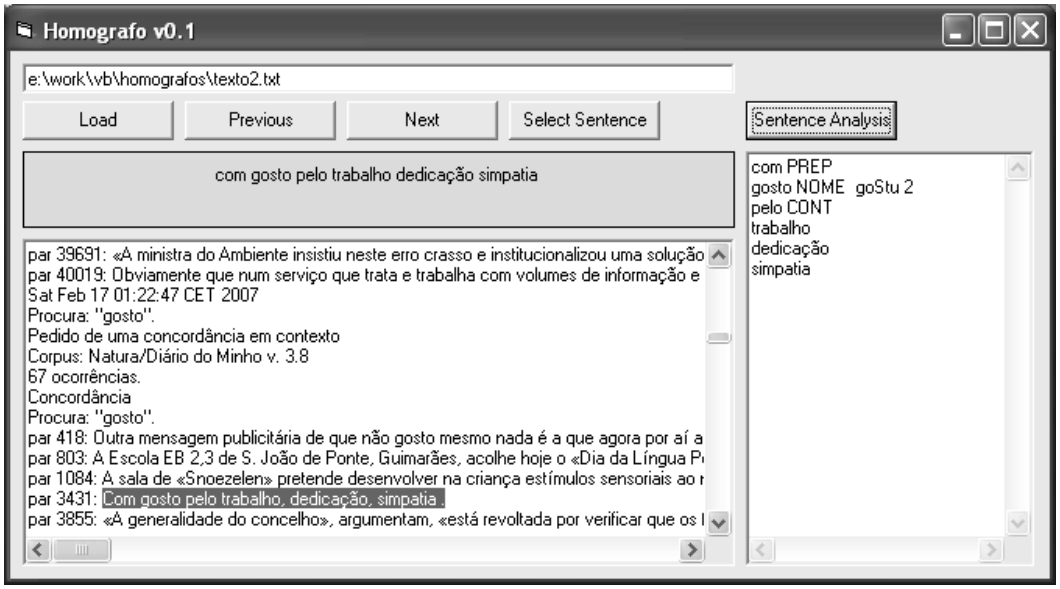

Figura 11: Interface do desambiguador de homógrafos.

Para teste do sistema, usou-se o corpus Natura-Diário do Minho[22], que contém excertos do jornal regional Diário do Minho e é constituído por 1738475 palavras.
Na Tabela 4, apresentam-se os resultados relativos à saída fonética do desambiguador de homógrafos. O teste do sistema revelou uma taxa de erro de 3,1%. Este valor é bastante animador quando pensamos que a percentagem de ocorrência dos homógrafos analisados neste corpus é apenas de 8430 em 1738475 palavras, o que dá uma percentagem de 0,48%.
As razões do bom desempenho dos algoritmos têm muitas vezes que ver, não com o facto de as perguntas cobrirem todos os contextos previstos, mas pelo facto de as respostas negativas poderem regressar à saída inicial, entendida como *default*.

tipo	homógrafo analisado	nº ocorrências no corpus	nº de erros	% de erros
1	'erro'	59	0	0,0
2	'gosto'	67	5	7,4
3	'rola	3	0	0,0
4	'colher'	3	0	0,0
5	'desses'	64	0	0,0
6	'fora'	primeiros 100 (de 252)	9	9,0
7	'seco'	4	0	0,0
8	'boto'	0	-	-
9	'este'	primeiros 100 (de 1946)	0	0,0
10	'leste'	39	0	0,0
11	'sobre'	primeiros 100 (de 2458)	0	0,0
12	'pegada'	0	-	-
13	'forma'	primeiros 100 (de 1154)	0	0,0
14	'cerca'	primeiros 100 (de 1327)	11	11,0
15	'besta'	0	-	-
16	'sede'	primeiros 100 (de 398)	8	8,0
17	'medo'	92	0	0,0
18	'termos'	primeiros 100 (de 523)	0	0,0
19	'cor'	34	0	0,0
20	'lobo'	1	0	0,0
21	'pregar'	6	0	0,0
Total	-	1072	33	3,1

Tabela 4: Resultados do sistema.

Outra razão tem que ver com o tipo de corpus usado nos testes que, pelo facto de ser jornalístico, apresenta um conjunto muito pouco variado de realizações de homógrafos, conduzindo quase sempre à saída mais provável. Esta mesma razão serve de justificação para o facto de não apresentarmos resultados para certos homógrafos (<boto>, <pegada> e <besta>), porque não ocorrem neste corpus. Por esta razão, foram usados três tipos de corpora durante a elaboração dos algoritmos, como foi exposto no ponto

[22] Este corpus encontra-se disponível para consulta em http://www.linguateca.pt/. Para mais informações sobre este projecto, consultar também: http://acdc.linguateca.pt/acesso/contabilizacao.html#minho.

3.1. De qualquer modo, é muito difícil encontrar vários tipos de texto num corpus. Em trabalhos futuros, testaremos os nossos algoritmos com corpus não jornalístico.
Os erros ocorrem na desambiguação de <gosto>, <fora>, <cerca> e <sede>. Todos os erros encontrados decorrem do aparecimento de contextos ou de combinações inesperados. Por exemplo, os erros na desambiguação de <cerca> decorrem de registos de horas que não tinham sido previstos (ex: "O fogo, que deflagrou **cerca** das 06h50 numa casa térrea...") ou outras co-ocorrências (ex. "O Papa não chegou a estar dez minutos na janela, **cerca** de metade do tempo das suas aparições dominicais."). Contextos deste tipo foram entretanto considerados.

5. Conclusões

Neste trabalho, desenvolveu-se um sistema de desambiguação de homógrafos heterófonos, baseado em regras linguísticas. Esta técnica provou ter um desempenho de 100% de acerto para 14 tipos de homógrafos.
Este sistema permite dar resposta ao problema da leitura dos homógrafos na conversão Texto-Fala do Português. O sistema é composto por 21 algoritmos que processam a desambiguação de 82 pares de homógrafos. A lista dos homógrafos recolhidos até ao momento de redacção deste trabalho também é apresentada.
O teste do sistema, efectuado com o corpus Natura-Diário do Minho, revelou uma taxa de acerto de 96,9%. Os erros apresentados serão tidos em consideração para futuros desenvolvimentos do sistema.
Está em curso um trabalho de recolha de homógrafos em Português do Brasil e em Galego, de forma a analisar a adaptabilidade dos nossos algoritmos a estas variedades linguísticas.
Como trabalho futuro, seria interessante avaliar a performance do nosso sistema ao nível do analisador morfossintáctico e semântico.
Prevemos ainda fazer uma comparação da técnica por nós utilizada com outras técnicas de desambiguação de homógrafos, como os métodos probabilísticos, nomeadamente as redes neuronais ou os modelos escondidos de Markov (Hidden Markov Models).

6. Bibliografia

Barbosa, F.; Ferrari, L.; Resende Jr., F. 2003. "A methodology to analize homographs for a Brazilian Portuguese TTS system". In *PROPOR'2003 - 6th Workshop on Computational Processing of the Portuguese Language*. Heidelberg: Springer-Verlag.
Bergström, M & Reis, N. 1997. *Prontuário ortográfico e guia da língua portuguesa*. Lisboa: Editorial Notícias.
Braga, D., Coelho, L.; Resende Jr., F. G. V. 2006. "A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese". In *VI International Telecommunications Symposium (ITS2006)*. Fortaleza-CE, Brasil.
Casteleiro, J. M. (coord.) 2001. *Dicionário da Língua Portuguesa Contemporânea da Academia das Ciências de Lisboa*. 2 vols. Lisboa: Editorial Verbo.
Cunha, C. & Cintra, L. 1992. *Nova gramática do português contemporâneo*. Lisboa: Sá da Costa.
Estrela, E.; Soares, M. A.; Leitão, M. J. 2004. *Saber escrever. Saber falar. Um guia completo para usar correctamente a língua portuguesa*. Lisboa: D. Quixote.

Ferrari, L ; Barbosa, F. ; 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*. Braga.
Huang, X.; Acero, A. and Hon, H.W. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. PTR, New Jersey: Prentice Hall.
Iriarte Sanromán, Álvaro. 2001. *A Unidade Lexicográfica. Palavras, colocações, frasemas, pragmatemas*. Centro de Estudos Humanísticos. Coleção Poliedro. Universidade do Minho.
Lyons, J. 1977. *Semantics.* 2 vols. Cambridge: Cambridge University Press.
Nogueira, R. Sá. 1994. *Dicionário de Verbos Portugueses Conjugados*. Lisboa: Clássica Editora.
Ramos, Emanuel (org.). s/d. *Os Lusíadas de Luís de Camões*. Porto: Porto Editora.
Ribeiro, R. Oliveira, L .C.; Trancoso, I. 2003 "Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese". In *PROPOR'200 3- 6th Workshop on Computational Processing of the Portuguese Language*. Springer-Verlag, Heidelberg, pp. 143-150.
Ribeiro, R.; Oliveira, L. C.; Trancoso, I. 2002. "Morphossyntactic Disambiguation for TTS Systems". *In Proc. of the 3rd Intl. Conf. on Language Resources and Evaluation*. Volume V. pp. 1427-1431. ELRA.
Seara, I.; Kafka, S. Klein, S.; 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". In *SBrT 2001 – XIX. Simpósio Brasileiro de Telecomunicações*. Fortaleza, Brasil.
Seara, I.; Kafka, S. Klein, S.; Seara, R. 2002. Alternância vocálica das formas verbais e nominais do Português Brasileiro para aplicação em conversão Texto-Fala. In *Revista da Sociedade Brasileira de Telecomunicações*. vol. 17, nº 1, pp. 79-85.
Villalva, A. 2003. "Formação de palavras: afixação", in *Mateus, M. H. M. (coord.) Gramática da Língua Portuguesa.* Lisboa: Caminho.
Yarowsky, D. 1996. "Homograph disambiguation in Text-to-Speech Synthesis". In *Progress in Speech Synthesis* (Jan van Santen, Richard Sproat, Joseph Olive, and Julia Hirschberg, editors), pp. 159-174, New York: Springer.